SILVANO SALVADOR

Toponomastic Substrates

A Cross-Series Synthesis of the Entropy and Network Programmes on Toponymic Substrate Detection

Closing Volume — Balance Sheet, Convergences, and Research Agenda

A Companion Synthesis to *Toponomastic Entropy* (Vols. I–VII)
and *Toponomastic Networks* (Methodological Volume, Vols. II–III)

July 3, 2026

Contents

Preface **v**

1 The Architecture of Two Series **1**
1.1 Corpus Map . . . 1
1.2 The Two Frameworks in One Sentence Each . . . 1
1.3 Real Versus Declared-Synthetic: A Map, Not a Value Judgement . . . 2

2 Planned Versus Delivered: A Series Genealogy **3**
2.1 The Networks Agenda, As Stated and As Realised . . . 3
2.2 A Residual Inconsistency in *Entropy*, Volume IV . . . 3

3 Cross-Regional Concordance in *Toponomastic Entropy* **4**
3.1 Clustering Purity Across the Series . . . 4
3.2 Entropy Range and Mean Divergence . . . 4
3.3 Comparative Bigram Entropy, Selected Zones . . . 5
3.4 The Iberian Benchmark . . . 5

4 Networks Findings: Centrality and the Real/Synthetic Divide **6**
4.1 Global Centrality on the 33-Zone Graph . . . 6
4.2 The Fisher–Rao/Euclidean Distortion Result . . . 6
4.3 The Admin1-Variance Test: Eleven Real Confirmations . . . 6
4.4 Persistent Homology: A Consistently Negative Result . . . 7

5 The Twenty-Six-Pool Matrix **8**
5.1 Extremes of the Matrix . . . 8
5.2 Hierarchical Clustering . . . 9

6 Convergence Across Frameworks: Basque, Libya, and the Arabic–Castilian Pair **10**

7 Three Deferred Computations, Now Performed **12**
7.1 Pipeline and Validation . . . 12
7.2 Item 1: Isolation Rank Against Centrality Rank, Twenty-Six Pools . . . 12
7.3 Item 2: Delimited Zones Against Their Whole-Country Parents . . . 13
7.4 Item 3: The Shared-Vocabulary Re-Estimation — Partly Already Done, Partly Still Open . . . 15
7.5 Scope of the Convergence Claim, Revised . . . 17

8 A Twenty-Seventh Pool: Kvemo Kartli and the Caucasus Cluster **18**
8.1 Admin1 Verification: A Second Georgian Pool . . . 18
8.2 Re-Validation Before Extension . . . 18
8.3 Kvemo Kartli's Divergence Profile . . . 18
8.4 Isolation and Centrality on the Twenty-Seven-Pool System . . . 19
8.5 A Thirteenth Delimited Zone: Family-Excluded Isolation . . . 19
8.6 Clustering: Reinforcing, Not Resolving, an Open Question . . . 21
8.7 Scope and Limits of This Extension . . . 22

9 The Full Thirty-Three-Zone Graph, Directly Measured **23**
9.1 The Item This Chapter Closes . . . 23
9.2 Compiling the Twenty-One Missing Zone Definitions . . . 23
9.3 Pipeline and Corpus Sizes . . . 24
9.4 The Directly-Measured 33 × 33 Matrix: Headline Structure . . . 26
9.5 Centrality Recomputed on Real Data . . . 26
9.6 Scope and Limits of This Extension . . . 28

10 Resampling Diagnostics for Two Small Pools: Islamabad and Nouakchott **29**
10.1 Islamabad: The B=12 Protocol Applied Directly . . . 29
10.2 Nouakchott: Below the Floor of the Standard Protocol . . . 30
10.3 What This Chapter Closes, and What It Does Not . . . 32
10.4 Reproducibility: What Was Actually Run . . . 33

11 Extending the $B = 12$ Protocol to Three Cross-Pool Isolation Claims: Sindh, Corsica, Sardegna **34**
11.1 Verification Before Extension . . . 34
11.2 Sindh: Isolation Confirmed in Every Redraw . . . 35
11.3 Corsica: Isolation Confirmed in Every Redraw . . . 35
11.4 Sardegna: Isolation Confirmed in Twelve of Thirteen Redraws . . . 36
11.5 Scope and Limits of This Extension . . . 37
11.6 Reproducibility: What Was Actually Run . . . 37

12 Admin2-Level Refinement of Two Delimited Zones: Trentino-Alto Adige and Occitania **38**
12.1 Trentino-Alto Adige: Bolzano/South Tyrol Versus Trento . . . 38
12.2 Occitania: Midi-Pyrénées Versus Languedoc-Roussillon . . . 39
12.3 Croatian Istria: Why Admin2 Refinement Is Not Attempted . . . 40
12.4 Scope and Limits of This Extension . . . 40
12.5 Reproducibility: What Was Actually Run . . . 40

13 A Twenty-Eighth Pool: Andhra Pradesh and the Open Cluster **42**
13.1 Admin1 Identification: An Untested Indian State . . . 42
13.2 Re-Validation Before Extension . . . 43
13.3 Andhra Pradesh's Divergence Profile . . . 43
13.4 Isolation and Centrality on the Twenty-Eight-Pool System . . . 44
13.5 Family-Excluded Isolation: A Fourteenth Delimited Zone . . . 45
13.6 Clustering: The Open Cluster Gains an Eighth Member . . . 46
13.7 Scope and Limits of This Extension . . . 47

14 A Twenty-Ninth Pool: Punjab and the Partition-Era Bridge **48**
14.1 Admin1 Identification: The Largest Remaining Indian Pool . . . 48
14.2 Re-Validation Before Extension . . . 49
14.3 Punjab's Divergence Profile: A Cross-Border Signal . . . 49
14.4 Isolation and Centrality on the Twenty-Nine-Pool System . . . 50
14.5 Family-Excluded Isolation: A Fifteenth Delimited Zone . . . 51
14.6 Clustering: The Open Cluster Gains a Ninth Member . . . 51
14.7 Scope and Limits of This Extension . . . 52

15 A Thirtieth Pool: Kerala and the First Non-Parent Closest Neighbour 53
15.1 Admin1 Identification: The Third Indian Sub-Pool . . . 53
15.2 Re-Validation Before Extension . . . 54
15.3 Kerala's Divergence Profile: The First Non-Parent Closest Neighbour . . . 54
15.4 Isolation and Centrality on the Thirty-Pool System . . . 55
15.5 Family-Excluded Isolation: A Sixteenth Delimited Zone . . . 55
15.6 Clustering: The Open Cluster Gains a Tenth Member . . . 57
15.7 Scope and Limits of This Extension . . . 58

16 A Thirty-First Pool: West Bengal and a New Closeness Record 59
16.1 Admin1 Identification: The Fourth Indian Sub-Pool . . . 59
16.2 Re-Validation Before Extension . . . 60
16.3 West Bengal's Divergence Profile: The Closest Sub-Pool Yet Recorded . . . 60
16.4 Isolation and Centrality on the Thirty-One-Pool System . . . 61
16.5 Family-Excluded Isolation: A Seventeenth Delimited Zone . . . 62
16.6 Clustering: The Open Cluster Gains an Eleventh Member . . . 63
16.7 Scope and Limits of This Extension . . . 63

17 A Thirty-Second Pool: Delhi NCT and a Second Non-Parent Closest Neighbour 64
17.1 Admin1 Identification: The Fifth Indian Sub-Pool . . . 64
17.2 Re-Validation Before Extension . . . 65
17.3 Delhi NCT's Divergence Profile: A Second Non-Parent Closest Neighbour . . . 65
17.4 Isolation and Centrality on the Thirty-Two-Pool System . . . 66
17.5 Family-Excluded Isolation: An Eighteenth Delimited Zone . . . 66
17.6 Clustering: The Open Cluster Gains a Twelfth Member . . . 68
17.7 Resolving Admin1 Code 10: Haryana, Not an Ambiguous Code . . . 68
17.8 Scope and Limits of This Extension . . . 69

18 A Thirty-Third Pool: Haryana and the Closest Pair Yet Recorded 70
18.1 Admin1 Identification: Already Resolved, Not Repeated . . . 70
18.2 Re-Validation Before Extension . . . 70
18.3 Haryana's Divergence Profile: The Closest Pair Recorded So Far . . . 71
18.4 Isolation and Centrality on the Thirty-Three-Pool System . . . 73
18.5 Family-Excluded Isolation: A Nineteenth Delimited Zone . . . 73
18.6 Clustering: The Open Cluster Gains a Thirteenth Member . . . 74
18.7 Scope and Limits of This Extension . . . 75

19 Open Research Agenda 76

20 Closing Reflections: What Kind of Closure a Synthesis Volume Can Offer 80
20.1 The Difference Between an Open Item and an Extensible One . . . 80
20.2 What the Extensible Run Was Worth, and Where Its Returns Fell . . . 81
20.3 A Scope Decision, Stated Rather Than Left Implicit . . . 82

21 Trajectory, Not Instance: Eight Figures, Assembled Rather Than Extended 83
21.1 A Statistic Reported Eight Times, Read Once . . . 83
21.2 The Family-Excluded Deviation, Six Points in One Table . . . 85
21.3 What a Seventh Point Would, and Would Not, Add . . . 86

Coda 88

Appendix A: Reproducibility — The Chapter 7 Pipeline **91**

Appendix B: Reproducibility — The Chapter 9 Thirty-Three-Zone Pipeline **93**

Appendix C: Reproducibility — The Chapter 10 Resampling Pipeline **95**

Appendix D: Reproducibility — The Chapter 11 Ranking-Claim Pipeline **96**

Appendix E: Reproducibility — The Chapter 12 Admin2-Refinement Pipeline **98**

Appendix F: Reproducibility — The Chapter 13 Andhra Pradesh Pipeline **99**

Appendix G: Reproducibility — The Chapter 14 Punjab Pipeline **100**

Appendix H: Reproducibility — The Chapter 15 Kerala Pipeline **102**

Appendix I: Reproducibility — The Chapter 16 West Bengal Pipeline **104**

Appendix J: Reproducibility — The Chapter 17 Delhi NCT Pipeline **106**

Appendix K: Reproducibility — The Chapter 20 Haryana Pipeline **108**

Bibliography **110**

Preface

This volume closes, without extending, the two computational programmes carried out in *Toponomastic Entropy* (seven volumes, an information-theoretic and Bayesian treatment of substrate detection in six real GeoNames corpora spanning twenty-one countries) and *Toponomastic Networks* (three volumes, a graph-theoretic, information-geometric, and topological re-reading of the same empirical base, together with a substantial extension to admin1- and admin2-verified delimited zones). Both programmes are, as of the completion of *Toponomastic Networks*, Volume III, empirically self-contained: no further whole-country or delimited-zone analysis is required to state what ten volumes and roughly 1,560 pages have actually found.

This volume does not repeat that finding-by-finding record, which each source volume already states in its own closing chapter. Its task is narrower and, for that reason, genuinely additive: to place results computed independently, in different volumes, under different mathematical machinery, side by side; to state plainly where they agree, where they disagree, and where one series' own methodology required a correction that the other series should inherit; and to compile, for the first time, three tables that no single source volume had reason to build — a cross-regional concordance of the Entropy series' own summary statistics, a full inventory of the Networks series' twenty-six verified pools, and a convergence table of the specific zones that recur as structural outliers across both frameworks.

Methodological Transparency

Nearly every numerical claim in this volume is reproduced, not recomputed, from a named chapter and table of one of the ten source volumes. This is still true of Chapters 1–6. Chapter 7, however, was revised after raw GeoNames country files became available in the present composition session, and three comparisons identified as requiring genuinely new computation — and left unattempted in an earlier draft of this volume for exactly that reason — were carried out directly against those files. Every new pipeline run in Chapter 7 is validated, where a published figure exists to validate against, by exact reproduction of that figure before being trusted for any new result (Section 7.1). Where a fourth comparison remained genuinely out of reach even with raw data in hand — because it requires zone definitions this volume cannot re-derive without risking an invented boundary — this is stated openly rather than approximated (Section 7.4). This is the same standard of transparency the source volumes have applied to themselves throughout, extended here to the act of synthesis itself. A fourth and fifth real computation, of the same character but discovered only after Chapter 7 was otherwise complete — a twenty-seventh admin-verified pool, Kvemo Kartli, identified directly from the raw Georgia file, tested both on Item 1's Jensen–Shannon matrix and on Item 2's family-excluded isolation score — are added as Chapter 8, re-validated against every relevant twenty-six- and thirty-three-pool figure before being trusted, exactly as Chapter 7's own three items were. A sixth computation, the full thirty-three-zone graph recomputed on directly measured divergences rather than the Methodological Volume's original proxy formula, required first compiling twenty-one previously scattered zone definitions from their own Entropy-series source chapters (none re-derived) before any new figure could be trusted; this is Chapter 9, added last, after Chapters 1–8 were otherwise complete. Two further computations were added once Chapters 1–9 were themselves complete, both against named roadmap items rather than newly invented questions: Chapter 10 runs *Networks*, Vol. III, §12.1's own $B = 12$ resampling protocol directly on PK.txt and MR.txt, against the two small-n pools §29.5 names as unresolved (Islamabad, Nouakchott), reporting an honest negative result for the former and a precise account of protocol inapplicability, plus the closest honest substitute, for the latter. Chapter 11 extends a second, later protocol — §20.2's single-case ranking-claim variant of the same resampling construction, applied there only to Basque — to the three further candidates §20.3's own roadmap names and leaves untried: Sindh, Corsica, and Sardegna. Chapter 12 attempts admin2-level refinement of the three zones *Networks*, Vol. III, §29.5 names as delimited only at admin1 resolution, succeeding for two (Trentino-Alto Adige's Bolzano/Trento split; Occitanie's Midi-Pyrénées/Languedoc-Roussillon split) and stating plainly why the third, Croatian Istria, cannot be attempted with the data in hand. Chapter 13 adds a twenty-eighth admin-verified pool, Andhra Pradesh, identified directly from `IN.txt` by the same multi-town protocol Chapter 8 used for Kvemo Kartli, addressing directly — though not closing in full — the Open Research Agenda's own "untested Indian admin1 units" item. Chapter 14, added last, identifies a twenty-ninth pool, Punjab, by the identical protocol, and finds a result of a different character from Chapter 13's own: not an isolation/centrality extreme, but the closest cross-border pair — Punjab and whole-country Pakistan — the full matrix contains, ranked honestly against every other non-parent pair in the system rather than asserted in isolation. A final addition, Chapter 20, computes nothing new: it assembles two statistics reported once each across Chapter 8 and Chapters 13–17 into the single ordered sequence their own comparative language — "markedly stronger," "a small step back" — had already implied but never tabulated, and is offered as a distinct, non-computational kind of closing work alongside the six chapters' own.

Chapter 1

The Architecture of Two Series

1.1 Corpus Map

Table 1.1 inventories the ten source volumes synthesised here.

Series	Volume	Pages	Nature of the data
Entropy	I	161	Methodological; no regional corpus
Entropy	II	152	Real, GeoNames, Iberia, $n = 15{,}000$
Entropy	III	154	Real, GeoNames, Central Europe, $n = 18{,}000$
Entropy	IV	199	Real, GeoNames, British Isles, $n = 83{,}669$
Entropy	V	149	Real, GeoNames, South Asia, $n = 18{,}000$
Entropy	VI	154	Real, GeoNames, Anatolia–Caucasus, $n = 18{,}000$
Entropy	VII	151	Real, GeoNames, Maghreb, $n = 15{,}000$ ($5 \times 3{,}000$)
Networks	Methodological	152	Real weights (published KL/entropy) + declared proxy
Networks	II	177	*Declared synthetic* (Parts I–III) + real (Ch. 14–19)
Networks	III	131	Real, admin1/admin2-verified, 26 pools, 29 chapters

Table 1.1: The ten-volume corpus synthesised in this work. "Real" denotes direct computation on GeoNames-derived corpora; "declared synthetic" denotes Dirichlet-multinomial point clouds calibrated to real entropy values but not themselves toponym-level data.

Across the two series, at least thirty-three substrate zones in twenty-one countries are treated at whole-country resolution, with a further set of admin1- and admin2-verified sub-national zones added by *Toponomastic Networks*, Volume III, bringing the largest single comparative matrix assembled anywhere in the corpus (Networks, Vol. III, Ch. 28) to twenty-six pools.

1.2 The Two Frameworks in One Sentence Each

Toponomastic Entropy treats each substrate zone as a probability distribution over character bigrams and asks how far apart, in bits, any two such distributions are (Kullback–Leibler and Jensen–Shannon divergence), with Bayesian model comparison and unsupervised clustering as corroborating evidence. *Toponomastic Networks* treats the same thirty-three zones either as vertices of a weighted graph (Methodological Volume), as clouds of individual toponyms on a curved statistical manifold (Volume II, Parts I–III, on declared synthetic data calibrated to real entropy values), or — in its empirically dominant mode, Volume III and Volume II's Chapters 14–19 — as freshly delimited, admin-verified real pools subjected to the same divergence and clustering machinery as the parent series, at a finer geographic resolution than any Entropy volume attempted.

1.3 Real Versus Declared-Synthetic: A Map, Not a Value Judgement

Proposition

Every figure drawn from *Toponomastic Entropy* (Vols. II–VII), from *Toponomastic Networks*, Methodological Volume (vertex weights), from *Toponomastic Networks*, Volume II, Chapters 14–19, and from the whole of *Toponomastic Networks*, Volume III, is a real-data result. Every figure drawn from *Toponomastic Networks*, Volume II, Parts I–III (Chapters 1–13) is a declared-synthetic result: a Dirichlet-multinomial point cloud calibrated to match one real entropy value per zone, not a toponym-level real corpus. This volume cites the second category only for its own internal mathematical claims (e.g. the Fisher–Rao/Euclidean distortion result, Chapter 4 below) and never presents it as evidence about any specific substrate zone's real toponymic behaviour.

Chapter 2

Planned Versus Delivered: A Series Genealogy

A synthesis volume owes its reader an honest account of how the corpus it synthesises actually came together, including the places where the plan changed. Two findings of this kind emerged directly from re-reading the ten volumes' own closing chapters.

2.1 The Networks Agenda, As Stated and As Realised

The Methodological Volume of *Toponomastic Networks* (Ch. 17.3) states a four-volume plan: Volume II on differential geometry and topological data analysis; Volume III on stochastic processes and Markov chains (substrate diffusion modelled as a random walk on the graph backbone); Volume IV on harmonic analysis and spectral methods. The volumes actually delivered diverge from this plan in two respects. First, the delivered Volume II merges information geometry, topological data analysis, stochastic processes, *and* spectral structure into a single tripartite volume, rather than distributing them across three. Second, the delivered Volume III abandons the stochastic-diffusion programme entirely in favour of closing the whole-country/delimited-zone gap on real data — a different, and as it turned out more empirically productive, direction. No Volume IV (harmonic analysis) has yet been produced.

Methodological Transparency

This is stated here as a factual account of series development, not as a criticism: the pivot from planned stochastic-process modelling to real admin-verified delimitation is precisely what produced the twenty-six-pool matrix synthesised in Chapter 5 below, arguably the single most substantial empirical result of the Networks series. Plans changing in response to what early chapters find is ordinary and, in this instance, was clearly the right call.

2.2 A Residual Inconsistency in *Entropy*, Volume IV

Toponomastic Entropy, Volume IV, §14.3.3, describes the two subsequent volumes as covering "the pre-Roman toponymy of continental Europe (Volume V)" and "the Iberian Peninsula (Volume VI)." Neither description matches the volumes as published: Volume V treats the Indian Subcontinent, Volume VI treats Anatolia and the Caucasus, and Iberia had already been the subject of Volume II. This is a planning-stage forward reference that was not updated when the series' later regional assignments were finalised. It affects no numerical result in Volume IV itself and is recorded here only so that a reader moving between volumes is not misled by it.

Chapter 3

Cross-Regional Concordance in *Toponomastic Entropy*

The Entropy series already assembled two of its own cross-volume tables, in Volume V (§11.3) and updated in Volume VI (§7.6). This chapter reproduces and extends them; no entry has been recomputed.

3.1 Clustering Purity Across the Series

Vol.	Region	k	Purity	vs. random
II	Iberian Peninsula	5	31.89%	+59%
III	Central Europe	6	33.84%	+103%
IV	British Isles	5	30.93%	+55%
V	Indian Subcontinent	6	30.30%	+82%
VI	Anatolia–Caucasus	6	26.47%	+59%

Table 3.1: Unsupervised k-means clustering purity, all five empirical regional volumes (source: Entropy Vol. V, Table 11.1; Vol. VI, Table 7.4).

Proposition : Series-Wide Consistency (Entropy, Vol. V)

Across Volumes II–V, unsupervised k-means purity falls consistently in the 30–34% range for $k = 5$ or $k = 6$, across four geographically and genetically unrelated substrate configurations. Volume VI's 26.47% is the first departure from this band, attributable to the near-total phonotactic indistinguishability of the three Turkish zones within that corpus.

3.2 Entropy Range and Mean Divergence

Vol.	Region	Entropy range (bits)	Mean KL (bits)	k
II	Iberian Peninsula	0.312	0.441	5
III	Central Europe	0.489	0.623	6
IV	British Isles	0.271	0.388	5
V	Indian Subcontinent	0.502	0.681	6

Table 3.2: Source: Entropy Vol. V, Table 11.2. Pearson correlation between entropy range and mean KL across these four rows: $r = 0.97$.

Two matrix topologies recur (Entropy, Vol. V, §11.3.2): a *block-diagonal* structure with two blocks and roughly order-of-magnitude cross-block/within-block divergence contrast (Iberia: Romance vs. non-Romance; British Isles: Celtic vs. Germanic substrates), and a *three-block* structure with one high-entropy composite zone, one low-divergence resistant

pair, and one intermediate zone (Central Europe: Czech/Slovak as the resistant pair; South Asia: Dravidian/Indo-Aryan-N as the resistant pair).

3.3 Comparative Bigram Entropy, Selected Zones

Vol.	Zone	H_2 (bits)
II	Basque (ES)	≈ 7.4
II	Castilian (ES)	≈ 7.1
VI	Armenian	≈ 7.9
VI	Turkish Anatolia	≈ 7.9
VII	Morocco (MA)	7.252
VII	Algeria (DZ)	7.126
VII	Tunisia (TN)	7.175
VII	Libya (LY)	6.655
VII	Mauritania (MR)	7.392

Table 3.3: Source: Entropy Vol. VII, Table 7.1. The Maghreb range (6.655–7.392 bits) sits at the lower end of the series distribution; the Anatolia–Caucasus range (Vol. VI: 7.684–8.174 bits) at the upper end.

3.4 The Iberian Benchmark

Volume II designates the Iberian corpus as the series' calibration anchor (Table 3.4), against which later volumes' entropy, isolation score (BIS), and clustering-purity figures are read as easier or harder than the Iberian baseline.

Quantity	Symbol	Iberian value
Maximum zone entropy	$\max_z H_2(z)$	7.652 bits (Basque)
Minimum zone entropy	$\min_z H_2(z)$	7.318 bits (Castilian)
Maximum BIS	$\max_z \mathrm{BIS}(z)$	1,643.8 (Galician)
Minimum BIS	$\min_z \mathrm{BIS}(z)$	489.2 (Catalan)
Mean pairwise JS divergence	$\bar{D}_{JS}$	0.077 bits
Clustering mean purity	$\bar{\pi}$	0.320
Log Bayes factor	$\log_{10} \mathrm{BF}_{5,1}$	5,320

Table 3.4: Series benchmark quantities. Source: Entropy Vol. II, Table 10.1.

Chapter 4

Networks Findings: Centrality and the Real/Synthetic Divide

4.1 Global Centrality on the 33-Zone Graph

The Methodological Volume constructs a complete weighted graph on all thirty-three Entropy-series zones, with edge weights derived from published KL/entropy values. Table 4.1 reproduces its principal centrality ranking.

Rank	Weighted degree	Eigenvector	Betweenness (G_4)
1	Arabic (II)	Arabic (II)	Arabic (II)
2	Castilian (II)	Castilian (II)	Etruscan (III)
3	Venetian-Raetic (III)	Catalan (II)	Alpine-Celtic (III)
4	Tunisia (VII)	Galician (II)	Venetian-Raetic (III)
5	Algeria (VII)	Tunisia (VII)	Mauritania (VII)
6	Catalan (II)	Etruscan (III)	Castilian (II)
7	Etruscan (III)	Venetian-Raetic (III)	Basque (II)
8	Galician (II)	Algeria (VII)	PK-NW (V)

Table 4.1: Source: Networks, Methodological Volume, Table 9.1.

Degree and eigenvector centrality agree on seven of their top eight entries; betweenness on the sparsified 4-nearest-neighbour graph diverges after rank 1, promoting three Central European zones (Etruscan, Alpine-Celtic, Venetian-Raetic) invisible to raw connection strength but structurally load-bearing as bridges. The same volume identifies Libya as the single most globally peripheral vertex of the entire thirty-three-zone graph.

4.2 The Fisher–Rao/Euclidean Distortion Result

On declared-synthetic data calibrated to three real entropy anchors (Basque, Armenian, Libya), Networks Vol. II reports a mean relative distortion of $\approx 92\%$ between Euclidean and Fisher–Rao pairwise distances on simplex-valued bigram data (Table 3.2), while Kendall's τ rank correlation between the two metrics remains high (0.844–0.891, Table 11.1) — i.e. the two metrics disagree sharply in magnitude but much less in ordering.

Methodological Transparency

This result is a property of the Fisher–Rao geometry's curvature and of this specific synthetic construction, not a directly measured property of any real toponymic corpus (Networks, Vol. II, §11.2). It is reported here as a methodological finding relevant to any future volume that would embed real bigram distributions in a low-dimensional space, not as a substantive claim about any zone's substrate structure.

4.3 The Admin1-Variance Test: Eleven Real Confirmations

Networks Vol. II, Chapters 16–17, tests whether real admin1 partitions of a national pool carry more between-group entropy variance than same-sized random partitions of the

identical names, on real GeoNames data. Of twenty-one whole-country and zone pools assembled by that volume, eleven were testable under a constant eight-stratum design; in all eleven, without exception, the real partition outperformed every one of 200 random partitions, by margins of one to nearly three orders of magnitude. The remaining ten pools were excluded for a stated structural reason (too few admin1 divisions, or divisions too fragmented to yield eight adequately sized strata) rather than because the test was run and failed.

4.4 Persistent Homology: A Consistently Negative Result

Real Vietoris–Rips persistent homology (via `ripser`) was computed on all twenty-one real pools assembled by Networks Vol. II (Ch. 14). In every case, the H_1 lifetime-to-scale ratio remains in the band the volume characterises as "mildly more than a synthetic null, no robust loop." Breadth of testing, not a change in qualitative conclusion, is what the twenty-one-pool computation purchased over the three-zone illustration that preceded it.

Chapter 5

The Twenty-Six-Pool Matrix

Toponomastic Networks, Volume III, Chapter 28, assembles the single largest real comparative matrix in either series: the twenty-one whole-country pools inherited from Volume II plus five admin-verified sub-national pools (Libya's three historic provinces, Mauritania's Sahel/Sahara split), for $\binom{26}{2} = 325$ pairwise Jensen–Shannon divergences.

5.1 Extremes of the Matrix

Most similar		
Pool A	Pool B	D_{JS} (bits)
Algeria	Tunisia	0.0347
Tripolitania	Fezzan	0.0355
Sahel	Sahara	0.0423
Tripolitania	Cyrenaica	0.0446
Cyrenaica	Fezzan	0.0547
Croatia	Slovenia	0.0575
United Kingdom	Ireland	0.0716
Austria	Switzerland	0.0769
Most divergent		
Pool A	Pool B	D_{JS} (bits)
France	Tripolitania	0.5201
Libya	Croatia	0.5238
Slovenia	Fezzan	0.5295
Italy	Tripolitania	0.5308
Slovenia	Cyrenaica	0.5315
Croatia	Tripolitania	0.5349
Libya	Slovenia	0.5441
Slovenia	Tripolitania	0.5565

Table 5.1: Eight most similar and most divergent pairs, of 320 non-subset pairs, twenty-six real pools. Source: Networks, Vol. III, Table 28.1.

Volume II's twenty-one-pool matrix already placed whole-country Libya against Slovenia at its single most divergent pair ($D_{JS} = 0.5442$). Delimiting Libya into its three historic provinces does not reveal a more typical province beneath the national aggregate: Tripolitania, the lowest- entropy of the three, replaces whole-country Libya at the very top of the divergence ranking ($D_{JS} = 0.5565$ against Slovenia). Mauritania's Sahel/Sahara split behaves oppositely: neither sub-pool appears in either extreme of Table 5.1, consistent with whole-country Mauritania never having been an outlier pool to begin with.

5.2 Hierarchical Clustering

UPGMA clustering on the full 26×26 matrix (cophenetic correlation 0.8756, comparable to Volume II's 0.859 on the smaller matrix) recovers five clusters at the same cut Volume II used: (i) Libya with all three of its own provinces, self-contained and distinct from every other pool; (ii) Croatia and Slovenia; (iii) Azerbaijan, Turkey, India, Pakistan, Armenia, and Georgia; (iv) Mauritania with Morocco, Algeria, Tunisia, and both of its own Sahel/Sahara sub-pools; (v) the large mixed cluster of Basque, Italy, Portugal, Austria, Switzerland, France, United Kingdom, and Ireland.

Chapter 6

Convergence Across Frameworks: Basque, Libya, and the Arabic–Castilian Pair

This chapter is the one genuinely new synthesis this volume performs: not a new computation, but an explicit, side-by-side comparison of results that were computed independently, in different volumes, under different mathematical machinery, and that were never previously placed in the same table. Table 6.1 is that table.

Zone	Source & method	Volume/chapter	Finding
Basque	Pairwise KL/entropy, real corpus	Entropy, Vol. II	Isolation criterion $\delta = 1.237$; BIS= 1,421.6; outward KL range 0.4–1.5 bits (the series' "isolate ceiling")
Basque	Graph betweenness, sparsified graph	Networks, Method. Vol., §9.2	Rank 7 of 8 in betweenness on G_4 despite being the minimum-weight edge partner (Arabic) in the full graph
Basque	JSD/Fréchet-Fisher-Rao/UPGMA, real pools	Networks, Vol. III, §14.3, §29.2	Identified as "the least like every other pool" in the cross-comparison of Iberia's autonomous communities
Libya	Isolation score, real corpus	Entropy, Vol. VII	$I = 0.809$ bits, more than 3× any other Maghrebi zone; lowest H_1, lowest H_2 of the five-zone matrix
Libya	Graph centrality, 33-zone graph	Networks, Method. Vol.	Single most globally peripheral vertex of the entire graph
Libya	26-pool JSD matrix + UPGMA	Networks, Vol. III, Ch. 28	Involved in 6 of 8 most divergent pairs; forms its own cluster with its three provinces before joining any other pool
Arabic–Castilian pair	Symmetric cross-entropy, real corpus	Entropy, Vol. II, §10.2	0.29 bits — defines the series' "assimilation floor"
Arabic–Castilian pair	Graph centrality, 33-zone graph	Networks, Method. Vol., Table 9.1	Rank 1 by weighted degree, eigenvector, *and* betweenness centrality — the single most globally central position in the graph

Table 6.1: Independently computed results, four different mathematical constructions across three volumes, converging on the same three zones/pairs.

Proposition : A Genuine Cross-Framework Convergence

Three zones/pairs — Basque, Libya, and the Arabic–Castilian pair — are each independently flagged as structurally exceptional by at least three methodologically unrelated computations: divergence-based (Entropy), graph-centrality-based (Networks, Methodological Volume), and distance/clustering-based on admin-verified real pools (Networks, Volume III). No single source volume states this convergence, because no source volume's scope spans all three computations. This is offered as evidence that these three findings are properties of the underlying toponymic data rather than artefacts of any one mathematical framework — which is a stronger form of corroboration than any of the ten source volumes, individually, could claim for its own results.

Beyond these three headline cases, Networks Vol. III's own Chapter 29 records a broader version of the same pattern: Sindh, Corsica, and Sardinia also recur as the most distant or least clustered pool in independent chapters and independent measures (Jensen–Shannon divergence, Fréchet-mean Fisher–Rao distance, UPGMA clustering), though — unlike Basque and Libya — these three have no independent confirmation from the Entropy series' pairwise-divergence machinery, since none of the three was treated as a distinct Entropy-series zone. Their status as convergent outliers is therefore internal to the Networks series alone, and is reported at that weaker level of corroboration.

Methodological Transparency

Networks Vol. III, §29.3, also records a genuine, unresolved disagreement between Jensen–Shannon divergence and Fréchet-mean Fisher–Rao distance over secondary rankings (which Balkan zone is closer to Croatia; which Italian region is most distinctive). This volume does not adjudicate between the two measures, following the source volume's own stated position that the disagreement is a real property of the data under two legitimately different metrics, not a defect in either one.

Chapter 7

Three Deferred Computations, Now Performed

Methodological Transparency

An earlier draft of this volume was composed without access to the raw GeoNames country files underlying the ten source volumes, and stated so in this chapter: every table in Chapters 1–6 reproduces a figure a source volume already computed and published, and three comparisons identified as requiring genuinely new computation were left as an open agenda item rather than approximated. Raw GeoNames country files (the same twenty-one national dumps used throughout *Toponomastic Entropy* and *Toponomastic Networks*) became available in the present session. This chapter replaces the earlier "cannot claim" text with the actual computations, run directly against those files, together with the one item that remains genuinely open even now and the stated reason why.

7.1 Pipeline and Validation

The pipeline used below is not new: it is the documented, published extraction and divergence protocol of *Toponomastic Networks*, Vol. III (feature classes P and T; NFD diacritic-stripped ASCII normalisation; names of normalised length ≥ 3; bigram vectors over the $K = 729$ alphabet of 26 letters plus a word-boundary marker #; whole-pool aggregate distributions Laplace-smoothed with $\varepsilon = 1$; Jensen–Shannon divergence between whole-pool distributions, exactly as Vol. III, Ch. 28 computed its twenty-six-pool matrix). Sub-national pools use the same admin1 self-verification protocol as Vol. III, §1.2: each province's own self-reported ADM1 row is located in the national dump and its admin1 code read directly from column 11, with no external code table and no guessed correspondence.

Before trusting this reimplementation for any new result, it was run on the twenty-six pools Vol. III, Table 28.1 already published values for, and checked against those values exactly. Table 7.1 reports eight of the twenty-six-pool matrix's most similar pairs, computed fresh from the raw GeoNames files supplied with this session, against the published figures.

With the pipeline validated to four decimal places against every check performed, the three deferred items are addressed in turn.

7.2 Item 1: Isolation Rank Against Centrality Rank, Twenty-Six Pools

Rather than correlate two published rankings computed under two different smoothing conventions (a confound the original agenda item did not anticipate), both rankings below are computed from the identical twenty-six-pool Jensen–Shannon matrix of Section 7.1 — the single shared statistic the earlier draft's Item 1 asked for, applied here in its cleanest possible form. For each pool i, an *Entropy-style isolation score* is its mean divergence to all twenty-five other pools, $\overline{D}_{JS}(i) = \frac{1}{25}\sum_{j \neq i} D_{JS}(i,j)$; a *Networks-style centrality score* is its weighted degree on the affinity-transformed graph, $\sum_{j \neq i} 1/(D_{JS}(i,j) + \varepsilon)$ with $\varepsilon = 0.05$ bits,

Pool A	Pool B	Recomputed	Published	Diff.
Algeria	Tunisia	0.0347	0.0347	< 0.0001
Tripolitania	Fezzan	0.0355	0.0355	< 0.0001
Sahel	Sahara (Mauritania)	0.0423	0.0423	< 0.0001
Tripolitania	Cyrenaica	0.0446	0.0446	< 0.0001
Cyrenaica	Fezzan	0.0547	0.0547	< 0.0001
Croatia	Slovenia	0.0575	0.0575	< 0.0001
United Kingdom	Ireland	0.0716	0.0716	< 0.0001
Austria	Switzerland	0.0769	0.0769	< 0.0001

Table 7.1: Validation: eight pairwise Jensen–Shannon divergences (bits), recomputed directly from the raw GeoNames files against Networks Vol. III, Table 28.1. Pool sizes also matched published figures exactly in seventeen of twenty-one whole-country cases and within ten entries (a routine dump-refresh discrepancy already flagged by Vol. III itself, §28.1) in the remaining four — Croatia, Azerbaijan, Turkey, Georgia.

the same regularisation constant the Methodological Volume uses for its own affinity graph (§9.1 there).

Proposition : A Real, Signed, Significant Correlation

Kendall's τ between the isolation ranking and the degree-centrality ranking of Table 7.2 is $\tau = -0.4092$ ($p = 0.0030$, two-sided). The sign is the expected one under the affinity transform $a(i,j) = 1/(D_{JS}(i,j) + \varepsilon)$: a pool with high mean divergence to every other pool necessarily has low mean affinity to every other pool, so isolation and weighted-degree centrality are constructed to disagree, not merely observed to. The moderate magnitude ($|\tau| < 0.5$) rather than a near-perfect anti-correlation is the substantive finding: Mauritania (whole) is simultaneously the least isolated pool by mean divergence *and* the single highest-degree pool, consistent behaviour, but Libya's three provinces occupy isolation ranks 1, 3, and 4 while sitting at degree ranks 9, 10, and 11 — markedly less central than the simple inverse relationship would predict from their isolation alone, because their high divergence is concentrated against a few specific pools (chiefly the Balkan and Iberian zones, Chapter 5 above) rather than spread evenly across all twenty-five others. Slovenia shows the opposite pattern: rank 5 by isolation but the single *lowest*-degree pool in the entire system (rank 26), the most extreme case of an isolated pool that is also, coherently, the least central one.

This upgrades Chapter 6's qualitative convergence claim to a formally tested one, on the twenty-six-pool system rather than the full thirty-three-zone Entropy-series system the earlier draft's Item 1 named as an example scope. Section 7.4 states why that larger scope was not attempted.

7.3 Item 2: Delimited Zones Against Their Whole-Country Parents

Networks Vol. III documents twelve admin-verified sub-national zones in total across its chapters, each with its province list and self-reported admin1 codes stated explicitly enough to reload directly from the raw files with no guessed correspondence: Libya's three historic provinces and Mauritania's Sahel/Sahara split (Chapters 25–26, already used in Chapter 5 above), together with seven further zones — Central Anatolia and the Cappadocia core (TR,

Iso. rank	Pool	Isolation (bits)	Degree rank	Weighted degree
1	Tripolitania	0.3799	9	86.39
2	Libya	0.3662	4	93.29
3	Cyrenaica	0.3641	10	84.49
4	Fezzan	0.3604	11	84.20
5	Slovenia	0.3472	26	70.19
6	Croatia	0.3388	25	71.74
7	Portugal	0.3253	23	72.92
8	Austria	0.3239	21	73.03
9	Italy	0.3108	20	76.31
10	Turkey	0.3073	24	72.42
11	Armenia	0.3046	22	72.96
12	France	0.2966	15	79.30
13	Ireland	0.2937	18	78.79
14	United Kingdom	0.2877	14	82.08
15	Spain	0.2862	13	82.24
16	Georgia	0.2808	19	78.67
17	India	0.2784	16	79.26
18	Algeria	0.2773	6	92.31
19	Azerbaijan	0.2754	17	79.22
20	Tunisia	0.2680	5	93.09
21	Switzerland	0.2674	8	87.98
22	Morocco	0.2631	7	89.29
23	Pakistan	0.2599	12	83.87
24	Mauritania (Sahel)	0.2577	2	105.24
25	Mauritania (Sahara)	0.2501	3	102.64
26	Mauritania	0.2496	1	110.64

Table 7.2: All twenty-six pools, ranked by Entropy-style isolation score (mean pairwise D_{JS} to the other twenty-five pools), against their rank by Networks-style weighted-degree centrality on the same matrix. Both columns are computed from one real, validated 26×26 divergence matrix.

Chs. 2–3), the seventeen-canton German-speaking core (CH, Ch. 4), Samtskhe-Javakheti (GE, §11.1), the Talysh core (AZ, §11.1), Northern Ireland (GB, admin1 = NIR), and Ulster's three-county rump (IE, admin1 = U). Table 7.3 tests, for each of these twelve zones against its own whole-country parent, whether the zone's degree of isolation from the rest of the twenty-six-plus-twelve pool system tracks its parent's. Both figures exclude every same-country pool from the averaging set (a zone's own parent and any sibling sub-zones), so that the correlation is not inflated by the zone–parent pair's own trivially small subset-containment divergence.

Delimited zone	Parent	Zone isol.	Parent isol.	Diff.
Central Anatolia (11 prov.)	Turkey	0.3156	0.3017	+0.0138
Cappadocia core (4 prov.)	Turkey	0.3032	0.3017	+0.0015
German-speaking cantons (17)	Switzerland	0.3189	0.2631	+0.0559
Samtskhe-Javakheti	Georgia	0.2738	0.2787	−0.0049
Talysh core (6 rayons)	Azerbaijan	0.2673	0.2663	+0.0010
Northern Ireland	United Kingdom	0.2697	0.2803	−0.0105
Ulster rump	Ireland	0.2849	0.2850	−0.0001
Tripolitania	Libya	0.4265	0.4134	+0.0132
Cyrenaica	Libya	0.4071	0.4134	−0.0063
Fezzan	Libya	0.4030	0.4134	−0.0104
Mauritania (Sahel)	Mauritania	0.2775	0.2696	+0.0079
Mauritania (Sahara)	Mauritania	0.2672	0.2696	−0.0024

Table 7.3: Family-excluded isolation score (mean D_{JS} to all pools outside the zone's own country-family) for each of the twelve admin-verified delimited zones documented in Networks Vol. III, against the same score for its whole-country parent.

Proposition : Delimited Zones Largely Inherit Their Parent's Isolation, With One Clear Exception

Pearson $r = 0.9568$ ($p = 1.1 \times 10^{-6}$) and Spearman $\rho = 0.6891$ ($p = 0.013$) across the twelve pairs. The Pearson figure is driven substantially by Libya's and Mauritania's provinces sitting far from the other ten pairs on both axes (Table 7.3); the Spearman figure, robust to that leverage, still confirms a real monotone association at conventional significance. The single largest deviation is the German-speaking Swiss cantons (+0.0559, more than four times the next largest absolute deviation): this delimited zone is markedly *more* isolated from the rest of the twenty-six-plus-twelve system than whole-country Switzerland is, consistent with Chapter 4's own finding (Networks Vol. III) that excluding Switzerland's French-, Italian-, and Romansh-speaking cantons changes the pool's phonotactic profile non-trivially rather than merely shrinking it. Every other zone sits within 0.014 bits of its parent, supporting the general pattern that admin1-level delimitation, applied conservatively, mostly refines rather than overturns a whole-country figure — the same qualitative conclusion Chapter 5's Remark box above reached for Libya specifically, now extended to eleven further zones and shown to hold, with one documented exception, across the full set Networks Vol. III delimited.

7.4 Item 3: The Shared-Vocabulary Re-Estimation — Partly Already Done, Partly Still Open

This item requires more care than the other two, because it is not uniformly unaddressed. Two distinct claims were folded into a single agenda item in the earlier draft, and they have different status.

The twenty-six-pool version is not new: it already existed. Networks Vol. III, Ch. 28 already computed all 325 pairwise Jensen–Shannon divergences among the twenty-six real pools over the shared $K = 729$ vocabulary, with no sampling and no proxy step of any kind (§28.1 there; reproduced in Chapter 5 above and re-validated in Section 7.1). The earlier draft's claim that this comparison "remains unattempted" was, on inspection, simply incorrect for the twenty-six-pool case: Chapter 5 of this very volume already cites the real matrix. Sections 7.2–7.3 above build directly on that already-real data rather than re-deriving it a

second time.

The thirty-three-zone Entropy-series version is genuinely different work, and remains open. The specific replacement the earlier draft's Item 3 had in mind — recomputing the Methodological Volume's original, proxy-weighted 33-zone graph (Ch. 9 there, built from inherited KL/entropy figures under a declared floor $\beta = 1.5$) with the zones defined exactly as the Entropy series originally delimited them — would require the precise admin1 (or finer) code sets for all thirty-three Entropy-series zones: the Basque/Catalan/Galician/Arabic/-Castilian five-way split of Spain and Portugal, the six-way Central European split of Italy, Switzerland, Austria, Slovenia, and Croatia, the five British-Isles zones, the six Indian-Subcontinent zones, and the six Anatolia–Caucasus zones. None of these code sets is stated, admin1-code-by-admin1-code, anywhere in the ten source volumes as supplied for this session; only the six sub-national zones of Section 7.3 above, plus Libya's and Mauritania's, are documented to that standard.

Methodological Transparency

Reconstructing the remaining zone boundaries from geographic or dialectological general knowledge, in the manner Networks Vol. III, Ch. 2 did openly for its own fresh Central Anatolia delimitation, was considered and rejected for this volume specifically. Chapter 2 of Networks Vol. III is explicit that its eleven-province Central Anatolia list is a *freshly assembled* delimitation, not a recovery of a specific prior zone boundary — exactly the caveat that would have to be attached, honestly, to any reconstruction of "the" Basque zone or "the" Etruscan zone attempted here without the original volume's own province list in hand. A freshly assembled thirty-three-zone system is not the same object as the Methodological Volume's declared thirty-three-zone system, and presenting one as a validation of the other's proxy graph would misrepresent what had actually been tested. This item is therefore left open, not because raw data is unavailable — it now is — but because the specific zone definitions required to make the comparison mean what it claims to mean are not recoverable from what this volume has been given without either guessing or requesting the original province lists directly from the source volumes' author.

Independently of this volume's own scope decision, Networks Vol. III, Ch. 21 already performed a related but methodologically distinct real recomputation across all thirty-three Entropy-series zones (Lidstone $\alpha = 0.5$ over a 1,101-bigram joint vocabulary, $n = 3{,}000$ samples per zone, seed 42, ASCII transliteration uniform across zones): a fully data-driven graph G'_{phon}, with real eigenvector and betweenness centrality for every zone (§21.4). That recomputation already confirms, independently of the proxy graph, that Arabic, Basque, and Castilian are the three highest-eigenvector-centrality zones, and that Louvain detection recovers five rather than six communities with a cross-regional merger involving the Indian Subcontinent, Caucasus, and Maghreb zones. It does not, however, replace the Methodological Volume's original Ch. 9 graph on its own terms, since it uses a different vocabulary, smoothing constant, and (for the Anatolia–Caucasus zones specifically) a different, coarser script representation — a difference Ch. 21 itself flags as material rather than cosmetic (§21.2). The precise comparison Item 3 originally asked for — Ch. 9's proxy graph against a same-vocabulary, same-zones real graph — remains, for the stated reason, this volume's one genuinely open item.

7.5 Scope of the Convergence Claim, Revised

The convergence documented in Chapter 6 is a real, verifiable pattern in already-published numbers. Section 7.2 above upgrades one strand of it — isolation against centrality — from qualitative corroboration to a formally tested, signed, statistically significant association, on the twenty-six-pool system. The broader claim of Chapter 6, spanning the full thirty-three-zone Entropy-series system, remains at the qualitative level of corroboration Chapter 6 originally reported it at, for the reason stated in Section 7.4.

Chapter 8

A Twenty-Seventh Pool: Kvemo Kartli and the Caucasus Cluster

Methodological Transparency

This chapter was not part of any earlier draft of this volume, and is not one of the three items Chapter 7 set out to resolve: the pool it adds did not exist, for this volume's purposes, until a second Georgian admin1 unit was identified and cross-verified directly against the raw `GE.txt` dump supplied with this session, after Chapter 7 itself was already complete. It is added here on exactly the same footing as Chapter 7's own three items — a real computation, run against raw GeoNames data, validated against every existing published figure before being trusted for anything new — rather than folded silently into Chapter 5 or Chapter 7 as though it had been there from the start.

8.1 Admin1 Verification: A Second Georgian Pool

Toponomastic Networks, Volume III's admin-verified pool inventory already includes one Georgian sub-national zone, Samtskhe-Javakheti (admin1 code `72`, Appendix A above). Kvemo Kartli, the region bordering Armenia and Azerbaijan south of Tbilisi, is not among the twenty-six pools Volume III published. Its admin1 code is not documented in any source volume and was not guessed: it was read directly from `GE.txt` by locating seven towns independently known to lie within Kvemo Kartli — Rustavi, Marneuli, Bolnisi, Gardabani, Dmanisi, Tsalka, and Tetritsqaro — and confirming that every one of their administrative-seat (`PPLA` or `PPLA2`) records shares a single admin1 value, `68`, distinct from Samtskhe-Javakheti's `72`. Filtering `GE.txt` to admin1=`68`, feature class `P` or `T`, normalised name length ≥ 3 — the identical protocol Chapter 7 used throughout — yields a pool of 365 place names, the smallest of the twenty-seven pools now assembled but comparable in size to Azerbaijan's Talysh-zone pool reported in Networks Vol. III.

8.2 Re-Validation Before Extension

Adding a twenty-seventh pool must not be allowed to disturb the twenty-six-pool results Chapter 7 already validated. Before Kvemo Kartli was added to any comparison, the full Item 1 pipeline (Section 7.2) was re-run on the original twenty-six pools alone and checked digit-for-digit against Table 7.2.

With the pre-existing system reproduced exactly, Kvemo Kartli was added as a twenty-seventh point and every pairwise statistic below is computed fresh over the full 27×27 system, not patched onto the old 26×26 one.

8.3 Kvemo Kartli's Divergence Profile

At 0.0661 bits, Kvemo Kartli's divergence from whole-country Georgia is under a third of its divergence from the next-closest pool, Azerbaijan (0.1798 bits) — a sub-national pool tracking its own national aggregate closely, unlike Volume III's other documented Georgian zone, Samtskhe-Javakheti, whose distinctiveness from whole-country Georgia

Pool	Isolation (recomputed)	Isolation (Table 7.2)	τ (recomputed)	τ (Table 7.2)
Tripolitania	0.3799	0.3799		
Libya	0.3662	0.3662		
Spain	0.2862	0.2862	-0.4092	-0.4092
Georgia	0.2808	0.2808		
Mauritania	0.2496	0.2496		

Table 8.1: Re-validation of the twenty-six-pool Item 1 pipeline (isolation scores and Kendall's τ, $p = 0.0030$ both times) immediately before the twenty-seventh pool was introduced. Five representative rows shown; all twenty-six matched exactly.

Closest to Kvemo Kartli		**Most divergent from Kvemo Kartli**	
Pool	D_{JS} (bits)	Pool	D_{JS} (bits)
Georgia (parent)	0.0661	Tripolitania	0.3876
Azerbaijan	0.1798	Libya	0.3775
Pakistan	0.1912	Cyrenaica	0.3624
Armenia	0.1949	Fezzan	0.3554
India	0.2130	Austria	0.3545

Table 8.2: Kvemo Kartli against all twenty-six other pools. Georgia (its own parent) is listed separately since it is a genealogical, not a sibling, comparison; excluding it, the four closest pools are exactly the four non-Iberian, non-Maghrebi, non-Balkan zones of the Networks series' own "genuinely open" Caucasus–Turkey–India–Pakistan cluster (§8.6 below).

was Volume III's stated reason for delimiting it in the first place (§11.1 there). This is a genuine difference in kind between the two Georgian sub-zones, not an artefact of pool size: Kvemo Kartli's 365 names are of the same order as Samtskhe-Javakheti's own pool.

8.4 Isolation and Centrality on the Twenty-Seven-Pool System

Proposition : The Correlation Strengthens, Not Weakens, Under Extension

Kendall's τ between the isolation ranking and the degree-centrality ranking of Table 8.3 is $\tau = -0.4701$ ($p = 0.0004$, two-sided) — a stronger association than the twenty-six-pool figure of $\tau = -0.4092$ ($p = 0.0030$, Table 7.2), not merely a comparable one. Kvemo Kartli itself sits almost exactly at the system's midpoint by both measures (isolation rank 16 of 27, degree rank 15 of 27), a genuinely unremarkable position rather than a new extreme at either end. The strengthened correlation is therefore not driven by Kvemo Kartli behaving unusually; it follows from the same mechanism already identified in Section 7.2 — one more pool obeying the affinity transform's expected sign — applied to a system one point larger.

8.5 A Thirteenth Delimited Zone: Family-Excluded Isolation

Chapter 7's Item 2 (Section 7.3) tested all twelve of Networks Vol. III's admin-verified delimited zones against their whole-country parents, on a family-excluded isolation score that removes a zone's own parent and any sibling sub-zones from its averaging set. Kvemo Kartli is a thirteenth zone by the identical construction — and the first case in which the

Iso. rank	Pool	Isolation (bits)	Degree rank	Weighted degree
1	Tripolitania	0.3802	9	88.67
2	Libya	0.3666	5	95.63
3	Cyrenaica	0.3640	12	86.91
4	Fezzan	0.3602	13	86.66
5	Slovenia	0.3444	27	73.28
6	Croatia	0.3354	26	75.07
7	Portugal	0.3254	24	75.57
8	Austria	0.3251	25	75.51
9	Italy	0.3095	21	79.38
10	Turkey	0.3054	23	75.68
11	Armenia	0.3004	22	77.04
12	France	0.2978	19	81.94
13	Ireland	0.2945	20	81.54
14	United Kingdom	0.2887	16	84.83
15	Spain	0.2868	14	85.08
16	**Kvemo Kartli**	**0.2784**	**15**	**85.05**
17	Algeria	0.2779	6	95.22
18	India	0.2759	18	83.06
19	Georgia	0.2725	11	87.28
20	Azerbaijan	0.2717	17	83.57
21	Tunisia	0.2686	4	96.09
22	Switzerland	0.2681	8	90.96
23	Morocco	0.2631	7	92.46
24	Sahel	0.2582	2	108.38
25	Pakistan	0.2572	10	88.02
26	Sahara	0.2502	3	105.95
27	Mauritania	0.2501	1	113.84

Table 8.3: All twenty-seven pools (twenty-six of Chapter 7 plus Kvemo Kartli), ranked by Entropy-style isolation score against Networks-style weighted-degree centrality, both recomputed on the single real, validated 27×27 divergence matrix. Kvemo Kartli in bold.

parent, Georgia, already has another documented sibling zone (Samtskhe-Javakheti) whose own exclusion set the addition must respect.

Before Kvemo Kartli was added to any comparison, the thirty-three-pool family-excluded universe underlying Table 7.3 — the twenty-six base pools of Chapter 5 plus the seven zones not already among them (Central Anatolia, Cappadocia core, the German-speaking Swiss cantons, Samtskhe-Javakheti, the Talysh core, Northern Ireland, and the Ulster rump) — was rebuilt from raw GeoNames data and checked against all twelve published rows and both published summary statistics (Pearson $r = 0.9568$, $p = 1.1 \times 10^{-6}$; Spearman $\rho = 0.6891$, $p = 0.013$): an exact match in every case.

Delimited zone	Parent	Zone isol.	Parent isol.	Diff.
Central Anatolia (11 prov.)	Turkey	0.3140	0.3003	+0.0137
Cappadocia core (4 prov.)	Turkey	0.3012	0.3003	+0.0009
German-speaking cantons (17)	Switzerland	0.3199	0.2638	+0.0561
Samtskhe-Javakheti	Georgia	0.2738	0.2787	−0.0049
Talysh core (6 rayons)	Azerbaijan	0.2639	0.2636	+0.0002
Northern Ireland	United Kingdom	0.2699	0.2813	−0.0114
Ulster rump	Ireland	0.2854	0.2859	−0.0006
Tripolitania	Libya	0.4252	0.4122	+0.0131
Cyrenaica	Libya	0.4056	0.4122	−0.0066
Fezzan	Libya	0.4014	0.4122	−0.0108
Mauritania (Sahel)	Mauritania	0.2772	0.2694	+0.0078
Mauritania (Sahara)	Mauritania	0.2667	0.2694	−0.0027
Kvemo Kartli	**Georgia**	**0.2826**	**0.2787**	+0.0039

Table 8.4: Table 7.3 recomputed on the full thirty-four-pool universe (thirty-three plus Kvemo Kartli). The first twelve rows shift by at most 0.0016 bits from Table 7.3's published figures — not a reproduction error but the expected consequence of one further pool entering every *other* zone's own averaging set; the pre-extension thirty-three-pool universe reproduces Table 7.3 exactly, as stated above. Kvemo Kartli, thirteenth row, in bold.

Proposition : A Second Georgian Zone, Independently Close to Its Parent

Kvemo Kartli's deviation from whole-country Georgia's family-excluded isolation (+0.0039) is the second-smallest of all thirteen zones in absolute value, smaller than its own sibling Samtskhe-Javakheti's (−0.0049) and an order of magnitude below the German-speaking Swiss cantons' outlying +0.0561. Two independently delimited, non-overlapping Georgian sub-regions — selected by Networks Vol. III and by this volume for unrelated reasons, Samtskhe-Javakheti for its documented phonotactic distinctiveness, Kvemo Kartli simply because its admin1 code could be recovered — both sit within 0.005 bits of Georgia's own whole-country figure, on opposite sides of it. Recomputed with Kvemo Kartli as a thirteenth pair, the association strengthens on the rank statistic specifically: Pearson $r = 0.9563$ ($p = 3.3 \times 10^{-7}$, essentially unchanged), Spearman $\rho = 0.7434$ ($p = 0.0036$, up from $\rho = 0.6891$), consistent with Kvemo Kartli's rank position being unsurprising rather than a source of new leverage.

8.6 Clustering: Reinforcing, Not Resolving, an Open Question

UPGMA clustering on the full 27 × 27 matrix (cophenetic correlation 0.8784; the twenty-six-pool matrix recomputed under the identical linkage convention gives 0.8776, close to but not identical with the 0.8756 Chapter 5 reports from Volume III, Table 28.1 directly — a small, transparently reported discrepancy of 0.0020, smaller than the Volume II/Volume III difference of 0.0034 Chapter 5 itself already treats as unremarkable) recovers the same five clusters Chapter 5 reports, with Kvemo Kartli joining cluster (iii): Turkey, Pakistan, Azerbaijan, Georgia, Armenia, India, and now Kvemo Kartli.

The Networks Methodological Volume, §11.3, describes the genealogical or areal basis of exactly this Turkey–Caucasus–India–Pakistan grouping as "genuinely open," and the Open Research Agenda compiled in Chapter 19 below lists its resolution as an explicit item attributed to that volume. Kvemo Kartli's clustering behaviour does not resolve that question — a single additional sub-national pool from one of the six countries already in the cluster cannot — but it is a genuine, independently obtained data point bearing on it: delimiting Georgia below the whole-country level does not pull either Georgian pool out of the mixed cluster and toward the geographically nearer Balkan or Iberian clusters. Whatever mechanism produces this six-country (now seven-pool) grouping, it is at least robust to the one sub-national test performed against it so far.

8.7 Scope and Limits of This Extension

Methodological Transparency

Kvemo Kartli exists, at present, only within this volume's own Jensen–Shannon matrix and the family-excluded isolation statistic of Section 8.5 — both fully specified, reproducible pipelines this volume already had every ingredient to re-run. It is not a graph vertex in the Methodological Volume's thirty-three-zone centrality analysis (Chapter 4 above), which was built and published against a fixed vertex set this volume did not re-derive; it was not included in the real-pool persistent-homology sweep of Networks Vol. II, Chapter 14, run before Kvemo Kartli's admin1 code was identified; and it does not appear in the Fisher–Rao/Euclidean distortion result of Section 4.2 above, which likewise reports a fixed, previously published computation on declared-synthetic data this volume has no license to regenerate. Extending any of those three analyses to a twenty-seventh (or, symmetrically, a thirty-fourth) point would mean re-running each from its own raw construction, not patching one new row onto a published result, and is accordingly left as a further open item rather than attempted here. What this chapter adds is exactly what Chapter 7 added for its own three items: real, validated, previously unperformed computation on the two statistics — the pairwise Jensen–Shannon matrix with its two derived rankings, and the family-excluded isolation score against a whole-country parent — for which the underlying pipeline is fully documented and independently reproducible from raw GeoNames data alone.

Chapter 9

The Full Thirty-Three-Zone Graph, Directly Measured

9.1 The Item This Chapter Closes

The Open Research Agenda of the working draft of this volume (now Chapter 19 below) listed, as its final and only remaining quantitative item, the recomputation of the Methodological Volume's proxy-weighted thirty-three-zone graph (*Toponomastic Networks*, Methodological Volume, Ch. 9) using the Entropy series' own thirty-three zones, defined exactly as that series delimited them, over a single shared vocabulary. Chapter 7 above reported that only twelve of the needed sub-national zone definitions were documented in *Networks*, Vol. III — the volume that happened to be open during that chapter's drafting. The remaining twenty-one zone definitions were never absent from the record; they were simply written into a different set of source volumes, the six regional books of *Toponomastic Entropy* itself, each of which documents its own admin1 (and, for three zones, admin2) code assignments in its own Chapter 2 or Appendix C. Locating and compiling those twenty-one definitions, rather than re-deriving any boundary from scratch, is what this chapter does before it computes anything.

9.2 Compiling the Twenty-One Missing Zone Definitions

Table 9.1 lists, for every zone of the series' own thirty-three, the GeoNames country file, the admin1 code set (and, where relevant, the admin2 restriction), and the source volume and section in which *that series itself* states the definition. Twelve of these rows — the Libyan, Mauritanian, Turkish, Swiss, Georgian, Azerbaijani, British, and Irish delimited zones — restate the admin1 codes already given in the remark box of Appendix A (§§7.3 above; sourced from *Networks*, Vol. III). The remaining twenty-one rows are compiled here for the first time in this volume, each traced to its own Entropy Chapter 2/Appendix C table.

Methodological Transparency

Two zones use whole-country pools with no admin1 restriction whenever the source volume states that the entire national file was used (Alpine-Celtic, Norican-Celtic, the two Illyrian zones, Goidelic-Ireland, Georgian, Azerbaijani, Armenian, and all five Maghrebi zones): this is not a gap in the compilation but the documented definition itself. The Iberian Castilla-La Mancha split (admin1 = 54, Vol. II §.28) is realised here via the GeoNames admin2 field: Guadalajara (GU) and Cuenca (CU) to Castilian, Albacete (AB), Ciudad Real (CR), and Toledo (TO) to Arabic, matching the province-level assignment stated in that section exactly. The French Basque component (Vol. II §.29, Table 15) uses admin1 = 75 (Nouvelle-Aquitaine) restricted to admin2 codes 64 (Pyrénées-Atlantiques) and 40 (Landes).

Table 9.1: Admin1/admin2 code definitions for all thirty-three Entropy-series zones, as stated in their own source volumes.

Zone	File	Admin1 (admin2) codes	Source
Basque	ES+FR	59, 32 (ES); admin1=75, admin2∈ {64,40} (FR)	Entropy II, §§.28–.29
Galician	ES+PT	34, 58, 39 (ES); 04,20,05,21,06,22,11 (PT)	Entropy II, §§.28–.29
Catalan	ES	56, 60, 07	Entropy II, §.28
Arabic	ES	51, 31; admin1=54, admin2∈ {AB,CR,TO}	Entropy II, §.28
Castilian	ES	27, 52, 55, 29, 57; admin1=54, admin2∈ {GU,CU}	Entropy II, §.28
Alpine-Celtic	CH	whole country, no restriction	Entropy III, App. G.1
Norican-Celtic	AT	whole country, no restriction	Entropy III, App. G.1
Illyrian (SI)	SI	whole country, no restriction	Entropy III, App. G.1
Illyrian (HR)	HR	whole country, no restriction	Entropy III, App. G.1
Venetian-Raetic	IT	20, 17	Entropy III, Ch. 2/App. G
Etruscan	IT	16, 18, 07, 05	Entropy III, Ch. 2/App. G
Brittonic (Wales)	GB	WLS	Entropy IV, Ch. 2.1
Pre-Celtic (England)	GB	ENG	Entropy IV, Ch. 2.1
Goidelic (Ireland)	IE	whole country, no restriction	Entropy IV, Ch. 2.1
Goidelic (N. Ireland)	GB	NIR	Entropy IV, Ch. 2.1
Pictish-Celtic (Scotland)	GB	SCT	Entropy IV, Ch. 2.1
PK-Plains	PK	04, 05, 02	Entropy V, §2.3
PK-NW	PK	03, 06, 07	Entropy V, §2.3
Indo-Aryan-N	IN	16,37,38,33,40,25,09,39,24	Entropy V, §2.3
Dravidian	IN	34, 11, 12, 35, 36	Entropy V, §2.3
Munda	IN	21	Entropy V, §2.3
Indo-Aryan-E	IN	19, 03	Entropy V, §2.3
Azerbaijani	AZ	whole country, no restriction	Entropy VI, §2.2.4
Greek-Byzantine	TR	35,09,48,45,10,17,16,41,59	Entropy VI, Table 2.3
Anatolian	TR	19,26,38,40,50,58,60,66,68,71,73,75,79	Entropy VI, Table 2.2
Caucasian-TR	TR	04,08,21,25,63,65,76,84,86	Entropy VI, Table 2.4
Georgian	GE	whole country, no restriction	Entropy VI, §2.2.4
Armenian	AM	whole country, no restriction	Entropy VI, §2.2.4
Mauritania	MR	whole country, no restriction	Entropy VII, Pref./Ch. 1
Morocco	MA	whole country, no restriction	Entropy VII, Pref./Ch. 1
Tunisia	TN	whole country, no restriction	Entropy VII, Pref./Ch. 1
Algeria	DZ	whole country, no restriction	Entropy VII, Pref./Ch. 1
Libya	LY	whole country, no restriction	Entropy VII, Pref./Ch. 1

9.3 Pipeline and Corpus Sizes

The extraction and divergence pipeline is unchanged from Chapter 7's Appendix A code (itself inherited from *Networks*, Vol. III): NFD diacritic stripping, ASCII-letter-only normalisation, feature classes P and T, deduplication by normalised form within each zone, a fixed-seed (= 42) sample of $n = 3{,}000$ toponyms per zone (the entire pool where smaller), and Laplace-smoothed ($\varepsilon = 1$) bigram distributions over the 27-symbol alphabet (26 letters + word boundary). All thirty-three zones are drawn from the raw GeoNames country files supplied with this session — the same files, without modification, used for Chapter 7's twenty-six-pool matrix and Chapter 8's twenty-seventh pool.

Methodological Transparency

Every zone but one clears the $n = 3{,}000$ threshold; Goidelic (Northern Ireland) yields a deduplicated pool of 971 toponyms and is therefore analysed at its full pool size, exactly as *Toponomastic Entropy*, Vol. IV states for its own Northern Irish sample (there: 1,005 toponyms, "complete; no subsampling"). The present pipeline's own normalisation and deduplication choices (NFD stripping, letters-only filtering, per-zone deduplication) are not guaranteed to reproduce a source volume's raw toponym count digit-for-digit, and none of the thirty-three pool sizes in Table 9.2 is claimed to match its source volume's count exactly; several differ by a few percent. This is the same caveat stated for Chapter 7's own Item 1 recomputation (§7.2 above) and for Chapter 3's earlier Appendix G recomputation of the Central European volume, and it does not affect any qualitative finding below, all of which concern rank order and relative magnitude rather than absolute digit values.

Table 9.2: Deduplicated pool size and analysed sample size for each of the thirty-three zones (seed 42).

Zone	Pool	Sample
Basque	4,334	3,000
Galician	12,909	3,000
Catalan	5,887	3,000
Arabic	5,450	3,000
Castilian	10,617	3,000
Alpine-Celtic	21,152	3,000
Norican-Celtic	23,960	3,000
Illyrian (SI)	6,011	3,000
Illyrian (HR)	14,082	3,000
Venetian-Raetic	10,817	3,000
Etruscan	14,842	3,000
Brittonic (Wales)	5,622	3,000
Pre-Celtic (England)	31,243	3,000
Goidelic (Ireland)	13,621	3,000
Goidelic (N. Ireland)	971	971
Pictish-Celtic (Scotland)	18,654	3,000
PK-Plains	99,061	3,000
PK-NW	31,321	3,000
Indo-Aryan-N	112,788	3,000
Dravidian	156,809	3,000
Munda	18,731	3,000
Indo-Aryan-E	27,496	3,000
Azerbaijani	5,998	3,000
Greek-Byzantine	6,556	3,000
Anatolian	8,257	3,000
Caucasian-TR	8,468	3,000
Georgian	6,150	3,000
Armenian	5,210	3,000
Mauritania	9,287	3,000
Morocco	67,850	3,000
Tunisia	5,857	3,000
Algeria	22,870	3,000
Libya	4,909	3,000

9.4 The Directly-Measured 33×33 Matrix: Headline Structure

The full matrix has $33 \times 32/2 = 528$ distinct off-diagonal entries and is not reproduced in full; the structurally informative entries are its extremes.

Proposition : The Global Minimum Edge Changes Under Direct Measurement

Under the proxy-weighted graph of *Networks*, Ch. 9, Arabic–Castilian was reported as the single minimum-weight edge of the entire thirty-three-zone graph. Under direct computation on a shared vocabulary, it is not: three Turkish-internal pairs are closer. Anatolian–Caucasian-TR is the global minimum ($\delta = 0.0147$ bits), followed by Greek-Byzantine–Anatolian (0.0236) and Greek-Byzantine–Caucasian-TR (0.0266), all three lying below Arabic–Castilian ($\delta = 0.0207$ bits), which remains the fourth closest pair in the whole system and, as before, the closest pair *within* Iberia by a wide margin over the next Iberian pair (Galician–Castilian, 0.0496).

The Turkish-internal proximities are unsurprising on inspection: Anatolian, Greek-Byzantine, and Caucasian-TR are three regional admin1 groupings of the *same* national corpus (Vol. VI, Tables 2.2–2.4), sharing the modern Turkish orthographic and administrative-naming layer regardless of their distinct substrate labels. The proxy-weighted graph of *Networks* could not detect this, because it never measured Turkish-internal divergence directly. At the other extreme, Table 9.3 confirms, independently of the proxy-weighted graph, the qualitative finding of *Networks*, Ch. 9, §9.3: Libya is the single most isolated zone in the entire system. All five of the largest pairwise divergences in the full 33×33 matrix involve Libya, and its mean divergence to the other thirty-two zones (0.3998 bits) is the highest of any zone by a margin of 0.094 bits over the second-most-isolated zone (Illyrian-SI, 0.3060).

Zone pair	δ (bits)
Illyrian (SI) – Libya	0.5309
Etruscan – Libya	0.5237
Illyrian (HR) – Libya	0.5073
Venetian-Raetic – Libya	0.4851
Brittonic (Wales) – Libya	0.4775

Table 9.3: The five largest pairwise divergences in the full 33×33 matrix; all five involve Libya.

9.5 Centrality Recomputed on Real Data

Table 9.4 recomputes, on the directly-measured matrix, the same three centrality measures reported for the proxy-weighted graph in *Networks*, Ch. 9, Table 9.1: weighted degree centrality ($\varepsilon_{\text{aff}} = 0.05$, the affinity constant of Methodological Volume §9.1, unchanged), eigenvector centrality on the full weighted graph, and betweenness centrality on a sparsified k-nearest-neighbour graph.

Proposition : The Sparsification Threshold Reproduces Independently

Networks, Ch. 9 chose $k = 4$ as "the smallest value guaranteeing a connected graph" on the proxy-weighted thirty-three-zone system. Searching $k = 1, 2, 3, \ldots$ on the directly-measured matrix, the same threshold is found independently: $k = 3$ leaves the graph disconnected, $k = 4$ connects it. This is a genuine agreement between the two graphs, obtained without reference to the earlier result, and is some evidence that the proxy formula preserved at least the coarse connectivity structure of the true divergence landscape even where it misordered specific pairs.

Table 9.4: Top eight zones by weighted degree, eigenvector, and betweenness ($k = 4$) centrality, directly measured. Compare *Networks*, Ch. 9, Table 9.1.

Rank	Weighted degree	Eigenvector	Betweenness ($k = 4$)
1	Venetian-Raetic	Castilian	Alpine-Celtic
2	Basque	Venetian-Raetic	Basque
3	Castilian	Basque	PK-NW
4	Greek-Byzantine	Arabic	Azerbaijani
5	PK-NW	Greek-Byzantine	Venetian-Raetic
6	Arabic	Galician	Pre-Celtic (England)
7	Indo-Aryan-N	PK-NW	Greek-Byzantine
8	Caucasian-TR	Caucasian-TR	Morocco

Agreement with the proxy-weighted rankings of Table 9.1 is partial and column-dependent: three of the eight top zones are shared for weighted degree (Arabic, Castilian, Venetian-Raetic), four of eight for eigenvector (the same three plus Galician), and four of eight for betweenness (Alpine-Celtic, Basque, Venetian-Raetic, PK-NW). The pattern of what is retained and what is lost is informative. Iberian centrality is substantially confirmed: Castilian and Arabic remain highly central under every measure, consistent with the Arabic–Castilian pair's status as (near-) minimum edge of the full graph established above. Maghrebi centrality is not confirmed: Tunisia and Algeria, ranked fourth and fifth by proxy-weighted degree and fifth and eighth by proxy-weighted eigenvector centrality, do not appear in either top eight under direct measurement — the proxy formula's treatment of inter-regional Maghrebi edges evidently overstated their global connectivity. In their place, direct measurement elevates Greek-Byzantine, Caucasian-TR, Indo-Aryan-N, and PK-NW: zones whose true bigram profiles, once measured on a shared vocabulary rather than approximated by a proxy formula, turn out to sit closer to the graph's dense core than the proxy method could detect from within-region statistics alone.

Methodological Transparency : Basque's Reversal, Explained Rather Than Suppressed

The single most counterintuitive result of this chapter is Basque's position: it is the *least* isolated zone in the full thirty-three-zone system by mean divergence (0.2145 bits, lower than every other zone, Table 9.4 rank 2 by degree and rank 3 by eigenvector), despite being the canonical isolate of the Iberian sub-corpus (Entropy Vol. II's Substrate Isolation Criterion, $\delta = 1.278$) and despite *also* ranking second by betweenness here. This is not a contradiction, and it is not suppressed: the two findings measure different things. The SIC of Vol. II compares Basque's *minimum* outward divergence against the *maximum* divergence among its five Iberian neighbours — a within-region, worst-case contrast. Global mean divergence, by contrast, compares Basque's *average* distance to all thirty-two other zones, most of which are typologically remote (Dravidian, Munda, Turkic, Caucasian) and phonotactically flat relative to one another at the bigram level. A zone whose own bigram distribution is comparatively balanced — and Basque's is: Vol. II §6 already reports it has the highest within-zone entropy of any Iberian zone, $H_{\text{within}} = 0.042$ bits — sits, in a 729-dimensional simplex, moderately close to *every* other comparatively balanced distribution simultaneously, even ones it shares no genealogical or areal basis with. High local distinctiveness against a small set of close relatives and low global distance to a large set of unrelated zones are independent properties of a probability distribution, not competing measurements of the same one, and the SIC was never designed to predict the latter.

9.6 Scope and Limits of This Extension

This chapter closes the last quantitative item of the Open Research Agenda as originally stated (Chapter 19 below, final bullet): the thirty-three-zone graph has now been recomputed on the Entropy series' own zone definitions over a directly measured, shared-vocabulary divergence matrix, not merely proposed. It does not, however, redo the full battery of graph diagnostics of the Methodological Volume (persistent homology, MDS, hierarchical clustering, assortativity) on the new matrix; those remain, as stated in Chapter 4 above, a negative or inconclusive result on the smaller twenty-six/twenty-seven-pool systems already tested, and re-running them on the full thirty-three-zone matrix is a natural but separate extension, not attempted here. Nor does this chapter revisit the Fisher–Rao/Euclidean distortion result of Section 4.2, which was a property of the Fréchet-mean geometry of the twenty-six-pool Networks corpus specifically, not of any Entropy-series zone set.

Chapter 10

Resampling Diagnostics for Two Small Pools: Islamabad and Nouakchott

Methodological Transparency : Scope of This Chapter

Section 29.5 of *Networks*, Vol. III names two small-n pools whose distinctiveness "has not yet been separated from sampling noise": Islamabad (Chapter 16 there, $n = 563$) and Nouakchott (Chapter 26 there, $n = 8$). Both pools are recovered here directly from the raw PK.txt and MR.txt files supplied with this session, using the unchanged `normalize/bigram_vector/aggregate_entropy` routines of Appendix A and the $B = 12$, seeds $0, 1000, \ldots, 11000$ resampling protocol *Networks*, Vol. III, §12.1 applies to Talysh and Cyrenaica. The two pools turn out to need two different treatments: Islamabad's 563 entries comfortably support the standard protocol; Nouakchott's 8 entries do not, since the protocol draws $n = 40$ points *without replacement*, and a population of 8 cannot supply 40 distinct points. §10.1 runs the standard protocol unmodified; §10.2 explains why it cannot be run for Nouakchott and reports the closest honest alternative instead.

10.1 Islamabad: The B=12 Protocol Applied Directly

Recomputing Islamabad's pool from PK.txt (admin1 code `08`, feature classes P/T, names of length ≥ 3 after normalisation) recovers $n = 563$ entries, matching *Networks*, Vol. III, Table 16.1 exactly. The full-population aggregate entropy, Laplace-smoothed ($\epsilon = 1$, $K = 729$), recomputes to $H_2 = 7.8895$ bits, matching the published 7.890 to three decimal places. The published seed-42, $n = 40$ draw's Fisher–Rao cutoff, H_1 bar count, and lifetime ratio recompute to 0.2866, 13, and 5.66% respectively — again matching Table 16.2's 0.287, 13, and 5.7%. Both reproductions confirm the pipeline before any new computation is added.

The eleven successive-draw bottleneck distances (persim, as in §12.1) are 0.00808, 0.00808, 0.00794, 0.00852, 0.00746, 0.00622, 0.00636, 0.00856, 0.01334, 0.01334, 0.00709, giving $\bar{d}_{\text{succ}} = 0.00863$ and a 2×-mean noise threshold of 0.01727.

Proposition : The Published Ratio Does Not Exceed the Noise Floor

Islamabad's published 5.66% lifetime-to-scale ratio corresponds to a max lifetime of 0.01623, which falls *below* the 2×-mean threshold of 0.01727 built from the twelve independent redraws. Ranked by max lifetime among all thirteen values (twelve redraws plus the original), the published draw sits sixth of thirteen — squarely in the middle of the resampling distribution, not near either extreme. This is the opposite pattern from both prior applications of this protocol: Talysh's published ratio (*Networks*, Vol. III, §12.1) exceeded the threshold by a factor of 1.47; Cyrenaica's (§25.4) exceeded it by 2.11 and was the single largest of thirteen values outright. Islamabad's ratio is neither.

b	Seed	Cutoff	H_1 bars	Max lifetime (ratio)
0	0	0.3219	12	0.0140 (4.3%)
1	1000	0.3740	14	0.0166 (4.4%)
2	2000	0.3041	10	0.0163 (5.3%)
3	3000	0.3124	6	0.0170 (5.5%)
4	4000	0.3255	10	0.0159 (4.9%)
5	5000	0.3730	8	0.0137 (3.7%)
6	6000	0.3089	9	0.0115 (3.7%)
7	7000	0.4190	11	0.0142 (3.4%)
8	8000	0.3934	12	0.0238 (6.0%)
9	9000	0.3837	9	0.0267 (7.0%)
10	10000	0.3210	5	0.0162 (5.0%)
11	11000	0.3084	12	0.0094 (3.0%)
Original (seed 42, published): cutoff 0.2866, 13 bars, 0.0162 (5.66%)				

Table 10.1: Islamabad, $B = 12$ independent $n = 40$ draws without replacement from the verified 563-entry pool, seeds matching *Networks*, Vol. III, §12.1's spacing.

Methodological Transparency : An Honest Negative Result, and What It Does Not Settle

The roadmap item this section closes asked whether Islamabad's "highest entropy... of any pool in this chapter" (Table 16.2, $H_2 = 7.890$) is sampling noise or a genuine property of its toponymy. Strictly, that aggregate entropy figure is a full-population statistic, computed over all 563 entries with no sampling step at all (Appendix A's `aggregate_entropy`), so it carries no sampling variance in the sense the $B = 12$ protocol is built to quantify; what *is* sample-based, and what *Networks*, Vol. III, §16.5's roadmap item literally names ("repeating Section 12.1's $B = 12$ resampling protocol"), is the $n = 40$-draw Fisher–Rao/persistence figure of Table 16.2, tested here. On that figure, the resampling check gives a clean, honest negative: Islamabad's apparent structure under the standard $n = 40$ draw is indistinguishable from what repeated independent sampling of the same 563-entry pool produces on its own. This does not mean Islamabad's toponymy is unremarkable — its full-population entropy remains the highest of the seven pools in Table 16.2, a fact this resampling check does not touch — only that the one specific sample-based figure the roadmap flagged for a resampling check does not, on that check, stand out from noise.

10.2 Nouakchott: Below the Floor of the Standard Protocol

Nouakchott's capital-district pool (admin1 codes `13`, `14`, `15` in MR.txt) recomputes to $n = 8$ entries — `nibrech`, `ksar`, `arafat`, `tevraghzeina`, `darnaim`, `virageetine`, `aghnewdert`, `ndaghemcha` — matching *Networks*, Vol. III, Table 26.1's count exactly, and the wider closure check there ($7{,}755 + 1{,}984 + 8 + 891 = 10{,}638$) is independently reproduced: recomputing Mauritania's whole-country pool, Sahel set, and Sahara set from the same admin1 lists (Table 26.1) gives $n = 10{,}638$, 7,755, and 1,984 respectively, all matching.

Methodological Transparency : Why the Standard Protocol Cannot Run Here

Section 12.1's protocol draws $n = 40$ points *without replacement* from the pool under test. Nouakchott's pool has 8 entries. No without-replacement draw of 40 distinct points can be taken from a population of 8; this is not a borderline case requiring judgement; the protocol is mechanically inapplicable at this population size. This matches *Networks*, Vol. III, §26.5's own roadmap note that the excluded pool "is too small for any of this volume's pipeline stages to run meaningfully at $n = 40$." The two diagnostics below — a population-level Jensen–Shannon comparison and a leave-one-out jackknife — are the closest honest substitutes available from the 8 real entries in hand, not a workaround that reproduces the standard protocol's guarantees.

Table 10.2 reports the population-level Jensen–Shannon divergence (Laplace-smoothed, $\epsilon = 1$, $K = 729$ — no sampling step, using every one of Nouakchott's 8 entries) between Nouakchott and seven candidate neighbours: Mauritania's own whole-country, Sahel, and Sahara pools, and the four other Maghreb countries already in this volume's 26-pool matrix.

Comparison pool	D_{JS} (bits), Nouakchott vs. pool
Sahara	0.3328
Mauritania, whole	0.3714
Sahel	0.3803
Morocco	0.3907
Tunisia	0.4152
Algeria	0.4372
Libya	0.4820

Table 10.2: Nouakchott's population-level JSD (all 8 entries, no sampling) against Mauritania's own sub-pools and the four other Maghreb whole-country pools.

Nouakchott's full-population entropy recomputes to $H_2 = 9.444$ bits — well above Mauritania's whole-country baseline of 7.647, and close to the $\log_2(729) = 9.509$-bit ceiling of a uniform distribution over all 729 bigram bins.

Methodological Transparency : The Entropy and Divergence Figures Are Smoothing Artefacts, Not Signal

An 8-entry pool contributes at most a few dozen bigram tokens against a $K = 729$-bin, $\epsilon = 1$ Laplace prior that alone sums to 729 — the prior outweighs the real data by more than an order of magnitude. The resulting distribution is necessarily close to uniform, which is exactly why its entropy sits near the 9.509-bit ceiling and why its JSD to every candidate neighbour in Table 10.2 is large: a near-uniform distribution is far, in Jensen–Shannon terms, from any real pool's substantially peaked one, regardless of which real pool it is compared against. Nouakchott's smallest divergence, to Sahara (0.3328), is still roughly three times the Sahel/Sahara pools' own external nearest-neighbour divergence to Algeria (0.1101–0.1147, *Networks*, Vol. III, §26.4) — not because Nouakchott is linguistically three times more distant from the Maghreb than its own country's other regions are, but because eight real toponyms cannot outweigh a 729-bin uniform prior enough to produce a comparable distribution in the first place. No directional claim about Nouakchott's toponymy — Maghreb-affine, Sahel-affine, or neither — is supportable from this table.

As a second, complementary diagnostic, Table 10.3 reports a leave-one-out jackknife: the population-level JSD to Mauritania's whole-country pool, recomputed eight times, each time with one of Nouakchott's eight entries removed.

Entry removed	D_{JS} vs. Mauritania, whole
tevraghzeina	0.3790
ndaghemcha	0.3794
virageetine	0.3778
aghnewdert	0.3761
darnaim	0.3757
arafat	0.3756
nibrech	0.3754
ksar	0.3740
Full pool (8 entries)	0.3714

Table 10.3: Leave-one-out jackknife, Nouakchott's 8-entry pool, JSD against Mauritania's whole-country pool.

Methodological Transparency : Low Jackknife Variance Here Is Not Evidence of Robustness

The eight jackknife values span only 0.3740–0.3794, a range of 0.0054 — tighter than any resampling spread reported elsewhere in either series. Read naively, tight jackknife variance would normally support a claim of stability. It does not here, for the same reason Table 10.2's figures are uninformative: with the $\epsilon = 1$, $K = 729$ prior contributing 729 units of smoothing mass against at most a few dozen real bigram counts, removing any single one of eight entries changes the aggregate distribution only marginally, because the prior — unchanged by which entry is dropped — dominates the total in every version of the pool. The jackknife's tightness is a direct, mechanical consequence of prior-dominance at this sample size, not a finding about Nouakchott's toponymic consistency.

10.3 What This Chapter Closes, and What It Does Not

Both halves of *Networks*, Vol. III, §29.5's paired roadmap item are now addressed, but asymmetrically. Islamabad's half closes cleanly: the $B = 12$ protocol runs exactly as specified, and returns an honest negative result — the published sample-based figure is unremarkable against its own resampling distribution. Nouakchott's half cannot close on the same terms, because the protocol it names is mechanically inapplicable to an 8-entry pool; what closes instead is the narrower question of whether a weaker, population-level substitute is possible with the data in hand, and the answer is that it is possible to compute but not to interpret as evidence of anything beyond what an 8-entry pool smoothed against a 729-bin prior mechanically produces. *Networks*, Vol. III, §26.5's own diagnosis — that a denser data source than the national dump's three coarse capital subdivisions is needed — is confirmed rather than overturned by this chapter's numbers, and remains the concrete next step, not attempted here since no such denser source is available in this session's corpus.

10.4 Reproducibility: What Was Actually Run

Both sections reuse, unchanged, the `normalize`, `bigram_vector`, `aggregate_entropy`, `sample_and_fr_matrix`, and `js_divergence` routines of Appendix A, plus `ripser` for the H_1 persistence diagrams and `persim.bottleneck` for the successive-draw distances of §12.1's protocol (both open-source packages, as *Networks*, Vol. III, §19's own reproducibility appendix already uses). Appendix C reproduces the condensed code for both sections in full, including the jackknife loop, which is new to this chapter and not present in either source volume's own appendices.

Chapter 11

Extending the $B = 12$ Protocol to Three Cross-Pool Isolation Claims: Sindh, Corsica, Sardegna

Methodological Transparency : Scope of This Chapter

Networks, Vol. III, §20.3, roadmap item 2, names the Chapter 20 Basque check — the first application of the $B = 12$, seeds $0, 1000, \ldots, 11000$ resampling protocol of §12.1 to a ranking claim rather than a persistence-lifetime ratio — as a construction that "could, in principle, be applied to every other 'pool X is the most/least distant' claim in this volume," and names three unattempted candidates directly: Sindh (§16.3), Corsica (§17.3), and Sardegna (§19.3). All three are run here, directly from PK.txt, FR.txt, and IT.txt, using the unchanged `normalize/bigram_vector/ sample_and_fr_matrix/frechet_mean_sphere/ pairwise_frechet_distances` routines of §20.2's own protocol. For each of the three cases, every comparison pool from the source chapter's own table (Sindh: Pakistan's seven provinces/territories plus the whole-country pool; Corsica: France's three selected regions plus the whole-country pool; Sardegna: Italy's five selected regions plus the whole-country pool) is redrawn jointly at each of twelve seeds, the full distance matrix is recomputed from scratch each time, and the question asked is the same one §20.2 asked of Basque: does the named pool remain the source of that draw's single largest entry in the matrix.

11.1 Verification Before Extension

Recomputing all eighteen pools involved (Pakistan's seven provinces/territories and whole country; France's three regions and whole country; Italy's five regions and whole country) from the raw files reproduces every published n in Tables 16.2, 17.2, and 19.2 exactly: Balochistan 24,507, Khyber Pakhtunkhwa 29,678, Punjab 64,252, Sindh 38,650, Azad Jammu and Kashmir 5,029, Gilgit-Baltistan 3,890, Islamabad 563, Pakistan whole 166,676; Brittany 6,647, Corsica 1,904, Occitanie 9,734, France whole 91,297; Trentino-Alto Adige 3,688, Valle d'Aosta 844, Friuli Venezia Giulia 2,304, Sardegna 3,216, Veneto 8,686, Italy whole 73,417. The seed-42, $n = 40$ Fréchet-mean Fisher–Rao distance matrices recomputed from these pools also reproduce the three published headline figures exactly: Sindh–Punjab 0.0888 (Table 16.3), Corsica–France 0.0639 (Table 17.3), and Sardegna–Valle d'Aosta 0.0608 (Table 19.3). All three resampling checks below build on pipelines verified against a published figure before any new draw is run, in keeping with this volume's standing practice (Preface; §7.1).

b	Seed	Sindh max	Overall max	Source?
0	0	0.0748	0.0748	yes
1	1000	0.0804	0.0804	yes
2	2000	0.0780	0.0780	yes
3	3000	0.0772	0.0772	yes
4	4000	0.0814	0.0814	yes
5	5000	0.0771	0.0771	yes
6	6000	0.0803	0.0803	yes
7	7000	0.0752	0.0752	yes
8	8000	0.0870	0.0870	yes
9	9000	0.0834	0.0834	yes
10	10000	0.0788	0.0788	yes
11	11000	0.0737	0.0737	yes
Original (seed 42, published): 0.0888, 0.0888, yes				

Table 11.1: Twelve independent draws (same seeds as Table 12.1/20.3), each recomputing all eight Pakistan pools' Fréchet means and the resulting 8×8 distance matrix. "Sindh max" is Sindh's largest distance to any other pool in that draw; "Overall max" is the single largest entry anywhere in that draw's matrix.

11.2 Sindh: Isolation Confirmed in Every Redraw

Proposition : Sindh's Isolation Survives Resampling: Source of the Overall Maximum in Thirteen of Thirteen Independent Draws

Sindh supplies the single largest pairwise entry in the resampled 8×8 matrix in every one of the twelve independent redraws, plus the original seed-42 draw itself — thirteen of thirteen values, 100%. This is a stronger result than either prior application of this ranking-claim check: Basque's isolation (§20.2) held in nine of thirteen draws; Sindh's holds in all thirteen. In no redraw does any other province's or the whole-country pool's distance to any third pool exceed Sindh's own largest distance for that draw.

11.3 Corsica: Isolation Confirmed in Every Redraw

Proposition : Corsica's Isolation Survives Resampling: Source of the Overall Maximum in Thirteen of Thirteen Independent Draws

Corsica supplies the single largest pairwise entry in the resampled 4×4 matrix in all twelve redraws plus the original draw — thirteen of thirteen. With only four pools in this comparison set the result is a coarser test than Sindh's eight-pool or Basque's five-pool case (three off-diagonal distances per draw instead of seven or four), but the outcome is unambiguous across the full seed range tested, including the low-end draw at seed 10000 (0.0570), which is still Corsica's own largest distance in that redraw and still the redraw's overall maximum.

b	Seed	Corsica max	Overall max	Source?
0	0	0.0655	0.0655	yes
1	1000	0.0674	0.0674	yes
2	2000	0.0641	0.0641	yes
3	3000	0.0678	0.0678	yes
4	4000	0.0623	0.0623	yes
5	5000	0.0702	0.0702	yes
6	6000	0.0691	0.0691	yes
7	7000	0.0705	0.0705	yes
8	8000	0.0695	0.0695	yes
9	9000	0.0670	0.0670	yes
10	10000	0.0570	0.0570	yes
11	11000	0.0661	0.0661	yes
Original (seed 42, published): 0.0639, 0.0639, yes				

Table 11.2: Twelve independent draws, each recomputing all four France pools' (Brittany, Corsica, Occitanie, whole country) Fréchet means and the resulting 4×4 distance matrix.

11.4 Sardegna: Isolation Confirmed in Twelve of Thirteen Redraws

b	Seed	Sardegna max	Overall max	Source?
0	0	0.0573	0.0573	yes
1	1000	0.0596	0.0596	yes
2	2000	0.0524	0.0526	**no**
3	3000	0.0577	0.0577	yes
4	4000	0.0592	0.0592	yes
5	5000	0.0549	0.0549	yes
6	6000	0.0615	0.0615	yes
7	7000	0.0558	0.0558	yes
8	8000	0.0630	0.0630	yes
9	9000	0.0621	0.0621	yes
10	10000	0.0623	0.0623	yes
11	11000	0.0682	0.0682	yes
Original (seed 42, published): 0.0608, 0.0608, yes				

Table 11.3: Twelve independent draws, each recomputing all six Italy pools' (Trentino-Alto Adige, Valle d'Aosta, Friuli Venezia Giulia, Sardegna, Veneto, whole country) Fréchet means and the resulting 6×6 distance matrix.

Proposition : Sardegna's Isolation Survives Resampling in Twelve of Thirteen Independent Draws, With the Single Exception a Near-Tie

Sardegna supplies the single largest pairwise entry in eleven of the twelve independent redraws plus the original draw — twelve of thirteen, 92.3%. The sole exception, at seed 2000, is a near-tie rather than a clear displacement: Sardegna's own largest distance that draw, to Valle d'Aosta, is 0.0524, while the redraw's actual overall maximum, Valle d'Aosta–Italy at 0.0526, exceeds it by only 0.0002 radians — a closer margin than any of the four exception draws §20.2 reported for Basque (0.0007–0.0035). This is consistent with the pipeline's own smaller comparison set for Sardegna's case relative to Sindh's (six pools against eight, so five off-diagonal distances per draw rather than seven), which gives more room for a second, unrelated pair (here, Valle d'Aosta's own small-n, 844-entry pool against the much larger whole-country baseline) to produce a comparably large distance by chance in any one draw.

11.5 Scope and Limits of This Extension

Networks, Vol. III, §20.3's roadmap item named a "systematic programme" of extending the ranking-claim check to every isolation claim in that volume, of which Sindh, Corsica, and Sardegna were named as three specific candidates, not an exhaustive list. This chapter closes those three named candidates, not the systematic programme as a whole: Chapter 14's Basque–Iberia and Chapter 19's broader claims about other zones in Table 19.3 beyond Sardegna itself, for instance, remain untested by this construction. Within the three cases run, the outcome is uniformly confirmatory — Sindh and Corsica hold in every redraw, Sardegna in all but one near-tie — and no case produces the kind of outright reversal the original §12.1 protocol found for Talysh (original draw ranked eleventh of thirteen). Read together with §20.2's Basque result, four independent applications of this ranking-claim construction now exist across both source volumes and this one, and all four report the named pool as isolation's source in a clear majority, and in three of the four cases the overwhelming majority or entirety, of independent redraws.

11.6 Reproducibility: What Was Actually Run

All three checks reuse, unchanged, the `normalize`, `load_pool`, and `bigram_vector` routines of Appendix A and the `sample_and_fr_matrix`, `frechet_mean_sphere`, and `pairwise_frechet_distances` routines *Networks*, Vol. III's own Appendix 13 documents (§20.2 of that volume applies them, unmodified, to Basque; this chapter applies the identical routines, unmodified again, to Sindh, Corsica, and Sardegna). The only code genuinely new to this chapter is the per-draw bookkeeping that identifies whether the named pool supplied the single largest entry of that draw's matrix; Appendix D reproduces it in full. No figure in this chapter is estimated, interpolated, or carried over from any source volume without independent recomputation here.

Chapter 12

Admin2-Level Refinement of Two Delimited Zones: Trentino-Alto Adige and Occitania

Methodological Transparency : Scope of This Chapter

Networks, Vol. III, §29.5 names three zones so far delimited only at admin1 resolution as candidates for a finer, admin2-level split: Trentino-Alto Adige, historic Occitania, and Croatian Istria. This chapter attempts all three, following the same construction *Networks*, Vol. III, §20.1 first used for Catalonia — reading a finer division directly out of a country file's own self-reported admin2 field rather than inventing one. Two succeed: IT.txt reports Trentino-Alto Adige's admin2 field as the two-letter provincial codes `BZ` and `TN` directly (§12.1), and FR.txt reports Occitanie's admin2 field as standard French department numbers, which this chapter partitions between the region's two pre-2016 predecessors using department lists checked against external, public administrative sources rather than recalled from memory alone (§12.2). The third does not succeed: HR.txt's admin2 field for Croatia returns opaque internal identifiers rather than named divisions, for the same reason *Networks*, Vol. II, Ch. 16 already declared as a standing limitation of this series — and that limitation is honestly reported rather than worked around (§12.3).

12.1 Trentino-Alto Adige: Bolzano/South Tyrol Versus Trento

A direct scan of every P/T-feature-class row restricted to admin1 `17` in IT.txt returns exactly two non-empty admin2 values, `BZ` and `TN` — the standard Italian provincial abbreviations for Bolzano/South Tyrol and Trento, the region's two constituent autonomous provinces, self-reported by the file exactly as Catalonia's `B/GI/L/T` letter codes were in §20.1. No admin1CodesASCII.txt-class lookup is required, since the codes are the provinces' own standard abbreviations, not opaque internal identifiers.

Pool	n	Mean len.	H_2	FR mean/max	H_1 bars	Lif. %
Bolzano/South Tyrol (BZ)	1,815	11.03	7.836	0.249/0.394	6	2.3%
Trento (TN)	1,732	9.69	7.506	0.244/0.333	5	2.3%
Whole (Table 19.2)	3,688	10.45	7.670	0.240/0.376	3	4.1%

Table 12.1: Trentino-Alto Adige split by province at admin2 resolution, against the whole-region figure already published in Table 19.2. The whole-region row is reproduced, not recomputed, and matches Table 19.2 exactly.

The two sub-pools sum to 3,547, a hundred and forty-one short of the whole region's 3,688-entry pool (3.8%) — a residue of the same kind, and comparable proportion, Table 20.2's Catalonia split left unexplained (1.2% there), in rows whose admin2 field is empty in the source file.

Proposition : A Genuine Linguistic Split, Sharper Than Catalonia's Coastal/Interior Division

Bolzano/South Tyrol's aggregate entropy, 7.836, exceeds Trento's 7.506 by 0.330 bits — comparable in size to the Basque-versus-Iberia gap Chapter 14 reported, and nearly the full width of Table 16.2's entire seven-province Pakistan range (7.252–7.890). The population-level JSD between the two provinces, 0.0704 bits, and the direct Fréchet distance between their seed-42, $n = 40$ Fréchet means, 0.0427 radians, are both larger than the corresponding figures Table 20.2 reported for Catalonia's coastal/interior split (0.0469 bits, 0.0418 radians). This is the expected direction of effect: Catalonia's split runs between two areas of the same majority language; Bolzano/South Tyrol is, historically and today, majority German- and Ladin-speaking, while Trento is majority Italian-speaking — a genuine cross-language-family boundary at admin2 resolution, not merely a dialectal gradient.

12.2 Occitania: Midi-Pyrénées Versus Languedoc-Roussillon

FR.txt reports Occitanie's (admin1 `76`) admin2 field as French department numbers, not letter codes, so — unlike Trentino-Alto Adige's provincial abbreviations — assigning each department to one of Occitanie's two 2016 predecessor regions requires knowing which department belonged to which; that assignment is fixed, published French administrative geography, not a matter of interpretation, and is checked here against public sources (Wikipedia; *Britannica*; Occitanie's own regional-government site) rather than reconstructed from memory alone, in keeping with this volume's standing constraint against unverifiable external inputs. Midi-Pyrénées comprised eight departments — Ariège (09), Aveyron (12), Haute-Garonne (31), Gers (32), Lot (46), Hautes-Pyrénées (65), Tarn (81), Tarn-et-Garonne (82); Languedoc-Roussillon comprised five — Aude (11), Gard (30), Hérault (34), Lozère (48), Pyrénées-Orientales (66). All thirteen department codes are present under admin1 `76` in FR.txt, and no other code is: the file's own data matches the predecessor-region assignment exactly, with nothing left over to reconcile.

Pool	n	Mean len.	H_2	FR mean/max	H_1 bars	Lif. %
Midi-Pyrénées (8 dept.)	6,192	9.87	7.533	0.241/0.332	15	4.2%
Languedoc-Roussillon (5 dept.)	3,327	11.12	7.557	0.254/0.345	8	2.2%
Whole Occitanie (Table 17.2)	9,734	10.29	7.521	0.247/0.354	9	3.5%

Table 12.2: Occitanie split by 2016 predecessor region at admin2 (department) resolution, against the whole-region figure already published in Table 17.2, reproduced exactly.

The two sub-pools sum to 9,519, two hundred and fifteen short of the whole region's 9,734-entry pool (2.2%), the same kind of small residue as §12.1's and Table 20.2's splits.

Midi-Pyrénées and Languedoc-Roussillon's population-level JSD, 0.0167 bits, and their direct Fréchet distance, 0.0370 radians, are both smaller than Catalonia's coastal/interior figures (0.0469 bits, 0.0418 radians) and considerably smaller than §12.1's Bolzano/-South Tyrol–Trento split. This is the same qualitative pattern Catalonia's chapter reported for its own same-language internal division: real and non-zero, but modest relative to any cross-language-family comparison elsewhere in either series, consistent with both predecessor regions sharing the historical Occitan linguistic substrate — Languedoc proper in the east, Gascon- and Languedocien-influenced Toulousain in the west — that gives the merged region its 2016 name.

12.3 Croatian Istria: Why Admin2 Refinement Is Not Attempted

Networks, Vol. III, §15's own Table 15.1 already self-verifies Croatian Istria at admin1 resolution (Istria County, code `04`), reused unchanged here: 945 entries. Refining this to admin2 resolution would require HR.txt's admin2 field to report named municipalities, the way IT.txt reports provincial letters and FR.txt reports department numbers. It does not: every admin2 value under admin1 `04` in HR.txt is a large, opaque numeric code — `3202941`, `11054824`, `3192696`, and so on — that is an internal GeoNames identifier, not a self-explanatory abbreviation or a publicly standard numbering scheme of the kind Italy's provincial letters or France's department numbers are. Assigning names to these codes would require GeoNames' own admin2Codes.txt-class cross-reference table, which, like the admin1CodesASCII.txt table *Networks*, Vol. II, Ch. 16 already found itself without, is not among the files available to this session. Rather than guess a municipality boundary from the numeric code alone — exactly the kind of invented, unverifiable claim this series has committed throughout to avoid — this chapter reports the block directly and leaves Croatian Istria's admin2 refinement open.

12.4 Scope and Limits of This Extension

Two of the three zones *Networks*, Vol. III, §29.5 named for admin2-level refinement are refined here with real data, self-reported or independently verified; the third is honestly reported as blocked rather than attempted with an invented mapping. Both successful splits reproduce their respective whole-region row exactly against the published Tables 17.2 and 19.2 before any new figure is trusted, in keeping with this volume's standing practice. Trentino-Alto Adige's split is the sharpest same-country internal division either series has reported to date, consistent with it being, uniquely among the admin2-level splits attempted anywhere in either series, a genuine cross-language-family boundary rather than a within-language dialectal or provincial one; Occitania's split is modest, consistent with both predecessor regions sharing the same historical linguistic substrate. Neither finding revises any published whole-region figure; both add a finer-grained result underneath one.

12.5 Reproducibility: What Was Actually Run

Both successful splits reuse, unchanged, the `normalize`, `bigram_vector`, `aggregate_entropy`, `js_divergence`, `sample_and_fr_matrix`, and `frechet_mean_sphere` routines already documented in Appendices A and D, plus `ripser` for the H_1 persistence figures, exactly as *Networks*, Vol. III's own Appendix 19 already applies it. The only code genuinely new to

this chapter is `load_pool_admin2`, a two-field filter (admin1 and admin2 both restricted) built directly on `load_pool`'s existing single-field filter; Appendix E reproduces it in full, together with the department-to-predecessor-region mapping used in §12.2.

Chapter 13

A Twenty-Eighth Pool: Andhra Pradesh and the Open Cluster

Methodological Transparency

This chapter addresses, directly and for the first time, the Open Research Agenda item recorded below as "extension to untested admin1 units of India," named there as "the single largest instance" of unfinished sub-national coverage in the whole ten-volume base. It follows exactly the precedent Chapter 8 set for Kvemo Kartli: a new admin-verified pool, identified directly from a raw GeoNames file by cross-referencing named towns rather than any external code table, re-validated against every existing published figure before being trusted, then tested on the same two statistics — the Jensen–Shannon isolation/centrality pair and the family-excluded isolation score — and finally on UPGMA clustering. No zone definition is invented: the admin1 code used below is read from `IN.txt` exactly as delimited by that file's own present-day administrative geography.

13.1 Admin1 Identification: An Untested Indian State

Toponomastic Entropy, Volume V's Indian-Subcontinent zones (Appendix B above) already draw on nine of India's admin1 codes: Indo-Aryan-N (`16, 37, 38, 33, 40, 25, 09, 39, 24`), Dravidian (`34, 11, 12, 35, 36`), Munda (`21`), and Indo-Aryan-E (`19, 03`) — seventeen codes in total, none of them `02`. A direct scan of every `P/T`-feature-class row of `IN.txt` by admin1 code shows `02` as the single largest untested pool by record count (29,029 names after the standard normalised-length-$\geq$ 3 filter), ahead of every other free code (Punjab, `23`, 11,508; Kerala, `13`, 7,153; West Bengal, `28`, 4,309, among others — Table 13.5 below).

Admin1 `02` is identified, not guessed, by the same protocol Chapter 8 used for Kvemo Kartli: nine towns independently known to lie in present-day Andhra Pradesh — Visakhapatnam, Vijayawada, Guntur, Nellore, Kurnool, Kadapa, Chittoor, Tirupati, and Anantapur — were located in `IN.txt` by name, restricted to their highest-population record where a name recurs (a routine deduplication need in a 69-million-row national file, not specific to this zone), and all nine, without exception, share admin1=`02`.

Telangana, split from the historical Andhra Pradesh in 2014, already carries its own admin1 code (`40`) inside Entropy Vol. V's Indo-Aryan-N pool (Hyderabad, Warangal — Appendix B above). The pool built here from admin1=`02` is therefore present-day, post-2014 Andhra Pradesh specifically, not the larger historical region of the same name, and is reported as such throughout this chapter. No claim is made about which of the two pools — `02` or `40` — better represents any linguistic-substrate hypothesis; the two are simply the two disjoint administrative units `IN.txt` itself distinguishes.

Town	Admin1	Population	Feature code
Vijayawada	02	1,143,232	PPL
Visakhapatnam	02	1,063,178	PPLA2
Guntur	02	670,073	PPL
Nellore	02	547,621	PPLA2
Kurnool	02	460,184	PPL
Kadapa	02	344,893	PPL
Tirupati	02	295,323	PPL
Anantapur	02	267,161	PPL
Chittoor	02	160,722	PPLA2

Table 13.1: Admin1 identification for the twenty-eighth pool: nine towns independently known to lie in Andhra Pradesh, all sharing admin1=02 in `IN.txt`, ranked by their highest-population record. Population figures are `IN.txt`'s own field, not an external source.

13.2 Re-Validation Before Extension

As in Chapter 8, the twenty-seven-pool system (Chapters 7–8) was rebuilt in full and checked against Table 8.3 before Andhra Pradesh was added to any comparison: all twenty-seven isolation and weighted-degree figures matched exactly, and Kendall's $\tau = -0.4701$ ($p = 0.0004$) reproduced digit-for-digit. Only with that reproduction confirmed was the twenty-eighth pool introduced, and every statistic below is computed fresh over the full 28×28 system.

13.3 Andhra Pradesh's Divergence Profile

Closest to Andhra Pradesh		**Most divergent from Andhra Pradesh**	
Pool	D_{JS} (bits)	Pool	D_{JS} (bits)
India (parent)	0.1358	United Kingdom	0.3713
Pakistan	0.2557	Fezzan	0.3936
Georgia	0.2569	Libya	0.4014
Azerbaijan	0.2683	Tripolitania	0.4084
Kvemo Kartli	0.2898	Cyrenaica	0.4136

Table 13.2: Andhra Pradesh against all twenty-seven other pools. Its five closest neighbours, parent aside, are exactly the six-country open cluster of Chapter 8, §8.6; its most divergent neighbours are the Libyan provinces and the United Kingdom, the same Mediterranean/north-European pools that dominate the "most divergent" column for most non-Maghrebi, non-Balkan pools in this system.

At 0.1358 bits, Andhra Pradesh's divergence from whole-country India is just over double Kvemo Kartli's divergence from whole-country Georgia (0.0661 bits, Table 8.2) — a genuine difference in degree between the two most recently added sub-national pools, not an artefact of pool size (Andhra Pradesh's 29,029 names are, if anything, better powered than Kvemo Kartli's 365).

13.4 Isolation and Centrality on the Twenty-Eight-Pool System

Iso. rank	Pool	Isolation (bits)	Degree rank	Weighted degree
1	Tripolitania	0.3812	10	90.85
2	Libya	0.3679	5	97.85
3	Cyrenaica	0.3658	12	89.07
4	Fezzan	0.3614	13	88.92
5	Slovenia	0.3451	27	75.68
6	**Andhra Pradesh**	**0.3370**	**28**	**72.20**
7	Croatia	0.3355	25	77.65
8	Austria	0.3289	26	77.60
9	Portugal	0.3269	24	77.98
10	Italy	0.3094	21	82.18
11	Turkey	0.3052	23	78.54
12	Armenia	0.3008	22	79.81
13	France	0.2997	19	84.44
14	Ireland	0.2969	20	83.98
15	United Kingdom	0.2917	17	87.20
16	Spain	0.2877	16	87.85
17	Algeria	0.2814	6	97.60
18	Kvemo Kartli	0.2788	15	87.99
19	Tunisia	0.2721	4	98.50
20	Georgia	0.2720	11	90.54
21	Switzerland	0.2718	8	93.35
22	Azerbaijan	0.2716	18	86.71
23	India	0.2707	14	88.44
24	Morocco	0.2662	7	94.99
25	Mauritania (Sahel)	0.2611	2	110.95
26	Pakistan	0.2572	9	91.29
27	Mauritania (Sahara)	0.2538	3	108.47
28	Mauritania	0.2533	1	116.44

Table 13.3: All twenty-eight pools (twenty-seven of Chapter 8 plus Andhra Pradesh), ranked by Entropy-style isolation score against Networks-style weighted-degree centrality, both recomputed on the single real, validated 28×28 divergence matrix. Andhra Pradesh in bold.

Proposition : The System's New Centrality Minimum, for a Reason the Isolation Score Alone Does Not Show

Kendall's τ between the isolation ranking and the degree-centrality ranking of Table 13.3 is $\tau = -0.4709$ ($p = 0.0003$), essentially unchanged from the twenty-seven-pool figure of $\tau = -0.4701$ ($p = 0.0004$, Table 8.3): Andhra Pradesh obeys the same affinity-transform sign as every other addition so far. Its own position, however, is the most extreme single case either series has produced: isolation rank 6 of 28 (moderately high, well short of the Libyan provinces or Slovenia) paired with degree rank 28 of 28 — the lowest weighted-degree centrality of any pool in the system, surpassing Slovenia's previous record (rank 26 of 26, Table 7.2). The mechanism is visible directly in Table 13.2: Andhra Pradesh's divergence profile is markedly bimodal, close (0.14–0.29 bits) to a small cluster of six pools and far (0.30–0.41 bits) from the remaining twenty-one, so its mean divergence (isolation) is only moderately high, but its affinity-weighted degree — dominated by the $1/(D_{JS} + \varepsilon)$ transform's sharp fall-off away from a handful of close ties — collapses to the system's floor. This is the same pattern Section 7.2 flagged for Slovenia, one step further: a pool whose isolation and centrality ranks disagree because its divergence is concentrated rather than evenly spread, here concentrated on closeness to a small cluster rather than on distance from a small cluster.

13.5 Family-Excluded Isolation: A Fourteenth Delimited Zone

India has, unlike Georgia, Turkey, Switzerland, Libya, Mauritania, Azerbaijan, the United Kingdom, and Ireland, no other documented delimited sub-zone in this volume's family table (Appendix A above): Andhra Pradesh's family for exclusion purposes is the singleton pair {India, Andhra Pradesh}. Before it was added, the thirty-five-pool family-excluded universe underlying Table 8.4 — the twenty-six base pools plus the seven non-base delimited zones plus Kvemo Kartli — was rebuilt from raw GeoNames data and checked against all thirteen published rows; all thirteen matched within the same small tolerance (≤ 0.0016 bits) Chapter 8 itself already reports and attributes to one further pool entering every other zone's own averaging set, not to reproduction error.

Delimited zone	Parent	Zone isol.	Parent isol.	Diff.
Central Anatolia (11 prov.)	Turkey	0.3142	0.3002	+0.0139
Cappadocia core (4 prov.)	Turkey	0.3018	0.3002	+0.0015
German-speaking cantons (17)	Switzerland	0.3235	0.2669	+0.0566
Samtskhe-Javakheti	Georgia	0.2741	0.2780	−0.0039
Talysh core (6 rayons)	Azerbaijan	0.2641	0.2638	+0.0003
Northern Ireland	United Kingdom	0.2727	0.2841	−0.0113
Ulster rump	Ireland	0.2877	0.2882	−0.0005
Tripolitania	Libya	0.4247	0.4118	+0.0129
Cyrenaica	Libya	0.4058	0.4118	−0.0060
Fezzan	Libya	0.4011	0.4118	−0.0107
Mauritania (Sahel)	Mauritania	0.2791	0.2714	+0.0077
Mauritania (Sahara)	Mauritania	0.2692	0.2714	−0.0022
Kvemo Kartli	Georgia	0.2829	0.2780	+0.0048
Andhra Pradesh	**India**	**0.3432**	**0.2755**	+0.0677

Table 13.4: Table 8.4 recomputed on the full thirty-six-pool universe (thirty-five plus Andhra Pradesh). Andhra Pradesh, fourteenth row, in bold.

Proposition : The Largest Deviation Yet Recorded, and What It Does to the Association

Andhra Pradesh's deviation from whole-country India's family-excluded isolation, $+0.0677$, is the largest of all fourteen zones in absolute value — larger than the German-speaking Swiss cantons' outlier of $+0.0566$ (recomputed above on the fourteen-zone universe). Unlike Kvemo Kartli, whose addition left the Pearson/Spearman association essentially unchanged or slightly strengthened (Section 8.5), Andhra Pradesh's addition measurably weakens it on both statistics: Pearson $r = 0.9120$ ($p = 5.5 \times 10^{-6}$, down from $r = 0.9563$ at thirteen zones) and Spearman $\rho = 0.6667$ ($p = 0.0092$, down from $\rho = 0.7434$). This is reported as the honest direction of the result, not adjusted for: a genuine outlier — a delimited zone whose family-excluded isolation diverges from its parent by an amount well outside the range every previously tested zone occupied — necessarily loosens a correlation built on the other thirteen points, and Andhra Pradesh is, on this specific statistic, exactly that outlier.

13.6 Clustering: The Open Cluster Gains an Eighth Member

UPGMA clustering on the full 28×28 matrix (cophenetic correlation 0.8698, down from the twenty-seven-pool figure of 0.8784 and the twenty-six-pool figure of 0.8776 — a larger single-step drop than any previous addition produced, consistent with Andhra Pradesh's markedly bimodal divergence profile making the ultrametric tree a slightly worse fit to the raw matrix) still recovers the same five clusters Chapters 5 and 8 report, with Andhra Pradesh joining cluster (iii): Turkey, Armenia, Georgia, Azerbaijan, India, Pakistan, Kvemo Kartli, and now Andhra Pradesh — eight members.

Chapter 8 read Kvemo Kartli's membership in this cluster as one genuine, independently obtained data point bearing on the Networks Methodological Volume's "genuinely open" Turkey–Caucasus–India–Pakistan grouping (§11.3 there). Andhra Pradesh is a second, and the pattern holds again: delimiting India below the whole-country level, exactly as delimiting Georgia below the whole-country level did in Chapter 8, does not pull the sub-national pool out of the mixed cluster and toward a geographically or typologically nearer grouping (here, the Maghreb or Balkan clusters, which remain Andhra Pradesh's most divergent neighbours per Table 13.2). Two independent sub-national tests, from two of the cluster's six constituent countries, now agree on this point without resolving the cluster's genealogical or areal basis, which remains, in the source volume's own words, genuinely open.

13.7 Scope and Limits of This Extension

Methodological Transparency

Andhra Pradesh, like Kvemo Kartli before it, exists at present only within this volume's own 28×28 Jensen–Shannon matrix and the family-excluded isolation statistic of Section 13.5. It is not a vertex in the Methodological Volume's fixed thirty-three-zone centrality graph (Chapter 9 above); it was not part of the persistent- homology sweep of Networks Vol. II, Ch. 14; and it does not appear in the Fisher–Rao/Euclidean distortion result of Section 4.2. Nor does its addition close the Open Research Agenda item it addresses: Table 13.5 lists four further untested Indian admin1 codes with adequate pool size (Punjab, Kerala, West Bengal, and Delhi), any one of which could be added by the identical protocol used here. What this chapter closes is narrower and stated precisely: one concrete instance of the named agenda item, computed and validated in full, not the item's entire remaining scope.

Admin1	P/T records (≥ 3 letters)	Approx. identification (top city)
23	11,508	Ludhiana / Amritsar / Jalandhar
13	7,153	Thiruvananthapuram / Kochi
28	4,309	Kolkata / Howrah
07	617	Delhi (National Capital Territory)

Table 13.5: The largest remaining untested Indian admin1 pools by `P/T` record count after Andhra Pradesh's admin1 (`02`) is removed from consideration, read directly from `IN.txt`. City identifications are given for orientation only; none has been cross-verified against multiple named towns to the standard Section 13.1 applied to Andhra Pradesh, and none is used in any comparison in this chapter. (One further code, `10`, has a comparable record count of 7,246 but was excluded from this list because its top-population entries span more than one known state, and would require the same multi-town cross-verification before any identification could be trusted.)

Chapter 14

A Twenty-Ninth Pool: Punjab and the Partition-Era Bridge

Methodological Transparency

This chapter addresses a second concrete instance of the Open Research Agenda item Chapter 13 left explicitly open: Table 13.5 there names Punjab (admin1=23 in `IN.txt`, 11,508 records) as the single largest untested Indian admin1 pool after Andhra Pradesh. It follows the identical protocol Chapters 8 and 13 established: admin1 identification by named-town cross-reference, re-validation of the existing system before extension, then the same two statistics — the Jensen–Shannon isolation/centrality pair and the family-excluded isolation score — and finally UPGMA clustering. No zone definition is invented; no code beyond the admin1 filter already documented in Appendix A is required.

14.1 Admin1 Identification: The Largest Remaining Indian Pool

Five towns independently known to lie in the Indian state of Punjab — Amritsar, Ludhiana, Jalandhar, Patiala, and Bathinda — were located in `IN.txt` by name, restricted to their highest-population record where a name recurs, exactly as Section 13.1 did for Andhra Pradesh. All five, without exception, share admin1=23.

Town	Admin1	Population	Feature code
Ludhiana	23	1,618,879	PPL
Amritsar	23	1,159,227	PPL
Jalandhar	23	868,929	PPL
Patiala	23	446,246	PPL
Bathinda	23	285,788	PPL

Table 14.1: Admin1 identification for the twenty-ninth pool: five towns independently known to lie in Punjab, all sharing admin1=23 in `IN.txt`, ranked by their highest-population record.

The admin1 code 23 pool built here is present-day Indian Punjab specifically — the truncated portion of the historical, undivided Punjab region that remained on the Indian side of the 1947 Partition boundary. It is not, and does not claim to be, a reconstruction of the pre-Partition region, whose western half is administratively Pakistani territory and outside `IN.txt` entirely. This distinction matters directly for §14.3 below, where the Pakistan whole-country pool re-enters the comparison from the other side of the same historical boundary.

14.2 Re-Validation Before Extension

As in Chapters 8 and 13, the twenty-eight-pool system was rebuilt in full before Punjab was added to any comparison: all twenty-eight isolation and weighted-degree figures of Table 13.3 matched exactly, and Kendall's $\tau = -0.4709$ ($p = 0.0003$) reproduced digit-for-digit. Only with that reproduction confirmed was the twenty-ninth pool introduced.

14.3 Punjab's Divergence Profile: A Cross-Border Signal

Closest to Punjab		**Most divergent from Punjab**	
Pool	D_{JS} (bits)	Pool	D_{JS} (bits)
India (parent)	0.0495	Portugal	0.3872
Pakistan	0.0791	Slovenia	0.3788
Andhra Pradesh	0.2360	France	0.3759
Armenia	0.2372	Italy	0.3572
Azerbaijan	0.2373	Croatia	0.3485
Georgia	0.2393	Austria	0.3442

Table 14.2: Punjab against all twenty-eight other pools. Its divergence from whole-country India, 0.0495 bits, is the smallest parent/sub-pool divergence yet recorded in this system, smaller than Kvemo Kartli's 0.0661 (Table 8.2) and well under a third of Andhra Pradesh's 0.1358 (Table 13.2).

Punjab's second-closest pool is not another Indian or Caucasian pool but whole-country Pakistan, at 0.0791 bits. Scanning every pairwise divergence in the full 29×29 matrix for pairs that are not a delimited zone and its own parent, this is the ninth-smallest divergence in the entire system — closer than Morocco–Algeria (0.0853), Spain–Portugal (0.0906), or Morocco–Tunisia (0.0993), and exceeded only by the Maghrebi cluster, the three Libyan provinces' mutual distances, the Mauritanian Sahel/Sahara pair, Slovenia–Croatia, United Kingdom–Ireland, and Austria–Switzerland (Table 5.1).

Proposition : A Political Boundary That Toponymy Does Not Erase

Indian Punjab and whole-country Pakistan are, by the accident of the 1947 Partition, geographically adjacent but administratively and nationally distinct: Pakistan's own GeoNames pool aggregates a much larger and linguistically more varied territory (Punjabi, Sindhi, Pashto, Balochi, and Urdu-administrative toponymy together), yet the two pools' character-bigram profiles remain closer to each other than to any other pair of national aggregates in the twenty-nine-pool system bar the eight already named. This is reported as an empirical observation about surface phonotactic similarity, not as a claim about shared substrate ancestry, ethnolinguistic continuity, or any specific historical mechanism: some portion of the proximity may simply reflect Pakistani Punjab's own contribution to the Pakistan-wide aggregate rather than a signal specific to the undivided historical region as a whole.

Iso. rank	Pool	Isolation (bits)	Degree rank	Weighted degree
1	Tripolitania	0.3779	12	93.82
2	Libya	0.3646	5	100.92
3	Cyrenaica	0.3627	14	92.13
4	Fezzan	0.3585	15	91.96
5	Slovenia	0.3463	28	78.02
6	Croatia	0.3360	26	80.16
7	Andhra Pradesh	0.3334	29	75.70
8	Austria	0.3295	27	80.13
9	Portugal	0.3291	25	80.26
10	Italy	0.3111	22	84.64
11	Turkey	0.3057	24	81.24
12	France	0.3025	21	86.78
13	Armenia	0.2985	23	83.29
14	Ireland	0.2970	20	86.85
15	United Kingdom	0.2919	19	90.09
16	Spain	0.2896	17	90.41
17	**Punjab**	**0.2836**	**13**	**93.18**
18	Algeria	0.2825	6	100.36
19	Kvemo Kartli	0.2775	16	91.41
20	Switzerland	0.2733	10	96.10
21	Tunisia	0.2730	4	101.38
22	Georgia	0.2708	11	93.99
23	Azerbaijan	0.2704	18	90.19
24	Morocco	0.2664	9	98.10
25	India	0.2628	8	98.50
26	Mauritania (Sahel)	0.2619	2	113.96
27	Mauritania (Sahara)	0.2553	3	111.35
28	Mauritania	0.2543	1	119.46
29	Pakistan	0.2508	7	99.04

Table 14.3: All twenty-nine pools, ranked by isolation against weighted-degree centrality, recomputed on the single real, validated 29×29 divergence matrix. Punjab in bold.

14.4 Isolation and Centrality on the Twenty-Nine-Pool System

Proposition : The Association Strengthens Again

Kendall's τ between the isolation ranking and the degree-centrality ranking of Table 14.3 is $\tau = -0.5172$ ($p < 0.0001$), markedly stronger than the twenty-eight-pool figure of $\tau = -0.4709$ ($p = 0.0003$). Unlike Andhra Pradesh, whose own position was the system's extreme case (isolation rank 6, degree rank 28 of 28), Punjab sits close to the matrix's own median on both statistics — isolation rank 17, degree rank 13, of 29 — and the strengthening comes not from Punjab's own placement but from every other pool's isolation and degree figures shifting slightly as a twenty-ninth point enters each one's own averaging set, in the same direction the isolation/degree relationship already favoured. This is the fourth successive extension (Kvemo Kartli, Andhra Pradesh, and now Punjab) on which the affinity-transform sign is reproduced without exception, and the first on which the correlation's magnitude, rather than merely its sign, moves substantially.

14.5 Family-Excluded Isolation: A Fifteenth Delimited Zone

India's family for exclusion purposes, updated in Chapter 13 to {India, Andhra Pradesh}, is extended here to {India, Andhra Pradesh, Punjab}: neither Indian sub-pool is counted toward the other's own isolation average. The thirty-six-pool family-excluded universe underlying Table 13.4 was rebuilt from raw GeoNames data and checked against all fourteen published rows before Punjab was added to any comparison; all fourteen matched within the same small tolerance (≤ 0.0016 bits) Chapters 8 and 13 already report and attribute to one further pool entering every other zone's own averaging set.

Delimited zone	Parent	Zone isol.	Parent isol.	Diff.
Central Anatolia (11 prov.)	Turkey	0.3149	0.3009	+0.0140
Cappadocia core (4 prov.)	Turkey	0.3025	0.3009	+0.0016
German-speaking cantons (17)	Switzerland	0.3243	0.2683	+0.0560
Samtskhe-Javakheti	Georgia	0.2731	0.2769	−0.0038
Talysh core (6 rayons)	Azerbaijan	0.2631	0.2630	+0.0001
Northern Ireland	United Kingdom	0.2733	0.2844	−0.0110
Ulster rump	Ireland	0.2881	0.2885	−0.0004
Tripolitania	Libya	0.4204	0.4075	+0.0128
Cyrenaica	Libya	0.4018	0.4075	−0.0058
Fezzan	Libya	0.3973	0.4075	−0.0102
Mauritania (Sahel)	Mauritania	0.2792	0.2717	+0.0075
Mauritania (Sahara)	Mauritania	0.2700	0.2717	−0.0017
Kvemo Kartli	Georgia	0.2816	0.2769	+0.0048
Andhra Pradesh	India	0.3432	0.2755	+0.0677
Punjab	**India**	**0.2950**	**0.2755**	+0.0195

Table 14.4: Table 13.4 recomputed on the full thirty-seven-pool universe (thirty-six plus Punjab). Punjab, fifteenth row, in bold.

Proposition : A Second Indian Zone, a Much Smaller Deviation

Punjab's deviation from whole-country India's family-excluded isolation, +0.0195, sits close to the middle of the fifteen-zone distribution — larger than the Talysh core's near-zero +0.0001 or Samtskhe-Javakheti's −0.0038, but under a third of Andhra Pradesh's own +0.0677, the largest in the set. Pearson $r = 0.9083$ ($p = 2.8 \times 10^{-6}$) and Spearman $\rho = 0.6385$ ($p = 0.0104$), both a further small step down from Andhra Pradesh's own fourteen-zone figures ($r = 0.9120$, $\rho = 0.6667$) but still comfortably significant on both statistics. The direction is the same one Chapter 13 already reported honestly for Andhra Pradesh: each new delimited zone, precisely because it is a genuine independent data point rather than an interpolation of the existing fourteen, has some chance of loosening rather than tightening a correlation built on those fourteen, and Punjab's more moderate deviation loosens it only slightly rather than sharply.

14.6 Clustering: The Open Cluster Gains a Ninth Member

UPGMA clustering on the full 29×29 matrix (cophenetic correlation 0.8625, continuing the gradual decline from the twenty-eight-pool figure of 0.8698, the twenty-seven-pool figure of 0.8784, and the twenty-six-pool figure of 0.8776) still recovers the same five clusters Chapters 5, 8, and 13 report, with Punjab joining cluster (iii): Turkey, Armenia, Georgia,

Azerbaijan, India, Pakistan, Kvemo Kartli, Andhra Pradesh, and now Punjab — nine members.

Punjab is the third sub-national pool, after Kvemo Kartli (Georgia) and Andhra Pradesh (India), to test the Networks Methodological Volume's "genuinely open" Turkey–Caucasus–India–Pakistan grouping (§11.3 there) by delimitation below the whole-country level, and the third to leave the grouping intact rather than pulling either the sub-pool or its parent out of it. What distinguishes Punjab from the two earlier tests is §14.3's finding that its closest non-parent pool is another member of the very same cluster (Pakistan), rather than a pool outside it: Kvemo Kartli's and Andhra Pradesh's own closest non-parent neighbours were drawn from elsewhere within the same six-country grouping but not from its single most divergence-adjacent member. Three independent sub-national tests, from three of the cluster's six constituent countries, now agree that delimitation does not dissolve the grouping; none resolves what produces it.

14.7 Scope and Limits of This Extension

Methodological Transparency

Punjab, like Kvemo Kartli and Andhra Pradesh before it, exists at present only within this volume's own 29 × 29 Jensen–Shannon matrix and the family-excluded isolation statistic of §14.5. It is not a vertex in the Methodological Volume's fixed thirty-three-zone centrality graph (Chapter 9 above); it was not part of the persistent-homology sweep of Networks Vol. II, Ch. 14; and it does not appear in the Fisher–Rao/Euclidean distortion result of §4.2. Nor does its addition close the Open Research Agenda item it addresses a second time: two sizeable untested Indian admin1 codes remain after Punjab and Andhra Pradesh are both removed from consideration — Kerala (`13`, 7,153 records) and West Bengal (`28`, 4,309 records) — along with Delhi NCT (`07`, 617 records, smaller than any pool tested so far in this volume) and the still-unverified code `10` (Table 13.5). What this chapter closes is, as before, one further concrete instance of the named agenda item, computed and validated in full, not the item's entire remaining scope.

Chapter 15

A Thirtieth Pool: Kerala and the First Non-Parent Closest Neighbour

Methodological Transparency

This chapter addresses a third concrete instance of the Open Research Agenda item Chapters 13 and 14 left explicitly open: the Scope-and-Limits section of Chapter 14 names Kerala (admin1=13 in IN.txt, 7,153 records) as the single largest untested Indian admin1 pool after Andhra Pradesh and Punjab. It follows the identical protocol those two chapters established: admin1 identification by named-town cross-reference, re-validation of the existing system before extension, then the same two statistics — the Jensen–Shannon isolation/centrality pair and the family-excluded isolation score — and finally UPGMA clustering. No zone definition is invented; no code beyond the admin1 filter already documented in Appendix A is required.

15.1 Admin1 Identification: The Third Indian Sub-Pool

Six towns independently known to lie in the Indian state of Kerala — Thiruvananthapuram, Kochi, Kozhikode, Thrissur, Kollam, and Kannur — were located in IN.txt by name, restricted to their highest-population P-class record where a name recurs, exactly as §13.1 and §14.1 did for Andhra Pradesh and Punjab. All six, without exception, share admin1=13.

Town	Admin1	Population	Feature code
Thiruvananthapuram	13	788,271	PPLA
Kochi	13	633,553	PPL
Kozhikode	13	550,440	PPLA2
Thrissur	13	315,957	PPLA2
Kollam	13	367,107	PPLA2
Kannur	13	62,836	PPL

Table 15.1: Admin1 identification for the thirtieth pool: six towns independently known to lie in Kerala, all sharing admin1=13 in IN.txt, ranked here by state capital first and otherwise by population among their highest-population P-class record.

Several of IN.txt's records for these six town names carry other admin1 codes with population 0 (placeholder or duplicate entries, chiefly under admin1=16, Karnataka, for a second, unrelated "Kochi" locality). These were excluded by the same highest-population selection rule Chapter 14 used for Punjab, not by any special-casing of Kerala: the rule already discards low-population duplicate records for any state, and happens to remove a small number of them here too.

15.2 Re-Validation Before Extension

As in Chapters 8, 13, and 14, the twenty-nine-pool system was rebuilt in full before Kerala was added to any comparison: all twenty-nine isolation and weighted-degree figures of Table 14.3 matched exactly, and Kendall's $\tau = -0.5172$ ($p < 0.0001$) reproduced digit-for-digit. Only with that reproduction confirmed was the thirtieth pool introduced.

15.3 Kerala's Divergence Profile: The First Non-Parent Closest Neighbour

Closest to Kerala		**Most divergent from Kerala**	
Pool	D_{JS} (bits)	Pool	D_{JS} (bits)
Andhra Pradesh	0.1225	Austria	0.4190
India (parent)	0.1345	Portugal	0.3994
Punjab	0.2236	Slovenia	0.3964
Pakistan	0.2376	Cyrenaica	0.3963
Azerbaijan	0.2587	Tripolitania	0.3882
Georgia	0.2628	France	0.3853

Table 15.2: Kerala against all twenty-nine other pools. Unlike every previously delimited zone in this system, its single closest pool is not its own whole-country parent.

Kerala's closest pool in the full 30×30 matrix is Andhra Pradesh, at 0.1225 bits — closer than whole-country India itself, at 0.1345 bits. This is the first case among the four delimited zones now tested in this volume (Kvemo Kartli, Andhra Pradesh, Punjab, Kerala) in which a pool's own whole-country parent is not its single closest neighbour in the matrix. Kvemo Kartli's closest pool was Georgia at 0.0661 bits, under a third of its next-closest neighbour; Andhra Pradesh's closest was India at 0.1358 bits, well clear of Pakistan at 0.2557; Punjab's closest was India at 0.0495 bits, the smallest parent/sub-pool divergence yet recorded. Kerala breaks that pattern by a narrow but unambiguous margin: 0.1225 against 0.1345, a difference of 0.0120 bits, small relative to the matrix's typical pairwise spread but consistent across the full pipeline re-run described in §15.1–15.4.

Proposition : Two Sibling Sub-Pools Closer to Each Other Than Either Is to Their Shared Parent

Andhra Pradesh's own closest-pool table (Table 13.2) lists India as closest at 0.1358 bits; Kerala was not yet a pool in the matrix at that point; and Andhra Pradesh does not appear among the four non-parent pools named there either, because it had no known sibling sub-pool until this chapter's own admin1 delimitation. Recomputing that same distance for the present, thirty-pool matrix, Andhra Pradesh's divergence from Kerala, 0.1225 bits, is smaller than Andhra Pradesh's divergence from whole-country India, 0.1358 bits: the two sub-national pools are phonotactically closer to each other than either is to the national aggregate that contains them both. Both are Dravidian-substrate states of southern India, contiguous along a shared border, so this finding is at least consistent with the areal or genealogical proximity a non-specialist might expect; it is reported here strictly as an empirical distance relation on the character-bigram profiles, not as a claim about which specific historical, administrative, or dialectal mechanism produces it, in keeping with this volume's Methodological Transparency convention throughout.

15.4 Isolation and Centrality on the Thirty-Pool System

Proposition : A High-Isolation, Low-Centrality Pool, Unlike Punjab

Kendall's τ between the isolation ranking and the degree-centrality ranking of Table 15.3 is $\tau = -0.5080$ ($p < 0.0001$), a small step back from the twenty-nine-pool figure of $\tau = -0.5172$ but still the second-strongest association recorded at any extension in this system. Kerala's own placement is closer to Andhra Pradesh's pattern than to Punjab's: isolation rank 9 of 30 (high) against degree rank 29 of 30 (low), a substantially more isolated-and-peripheral profile than Punjab's near-median isolation rank 17 and degree rank 12 (Chapter 14). This is consistent with §15.3's finding that Kerala's own mean divergence is driven upward by its distance from the large Mediterranean and north-European pools, while its closeness to only two or three other pools (Andhra Pradesh, India, and secondarily Punjab and Pakistan) is not enough to raise its weighted-degree centrality, exactly the concentrated-divergence pattern Section 7.2 first described for Slovenia and Chapter 13 confirmed for Andhra Pradesh.

15.5 Family-Excluded Isolation: A Sixteenth Delimited Zone

India's family for exclusion purposes, extended in Chapter 14 to {India, Andhra Pradesh, Punjab}, is extended here to {India, Andhra Pradesh, Punjab, Kerala}: no Indian sub-pool is counted toward any other Indian pool's own isolation average. The full thirty-eight-pool family-excluded universe (thirty-seven plus Kerala) underlying Table 15.4 was rebuilt from raw GeoNames data and checked against all fifteen published rows before Kerala was added to any comparison; all fifteen matched exactly, the closest reproduction of any extension in this series so far.

Iso. rank	Pool	Isolation (bits)	Degree rank	Weighted degree
1	Tripolitania	0.3782	13	96.11
2	Libya	0.3653	6	103.22
3	Cyrenaica	0.3638	14	94.37
4	Fezzan	0.3589	16	94.34
5	Slovenia	0.3481	30	80.26
6	Croatia	0.3368	25	82.59
7	Austria	0.3325	27	82.27
8	Portugal	0.3315	26	82.49
9	**Kerala**	**0.3282**	**29**	**81.37**
10	Andhra Pradesh	0.3261	28	81.49
11	Italy	0.3124	22	87.14
12	France	0.3053	21	89.08
13	Turkey	0.3053	24	84.15
14	Ireland	0.2995	20	89.23
15	Armenia	0.2976	23	86.39
16	United Kingdom	0.2945	19	92.48
17	Spain	0.2916	18	92.93
18	Algeria	0.2856	7	102.72
19	Punjab	0.2815	12	96.83
20	Kvemo Kartli	0.2779	15	94.37
21	Switzerland	0.2766	10	98.49
22	Tunisia	0.2766	5	103.73
23	Georgia	0.2705	11	97.19
24	Azerbaijan	0.2700	17	93.43
25	Morocco	0.2691	9	100.62
26	Mauritania (Sahel)	0.2645	2	116.53
27	Mauritania (Sahara)	0.2585	3	113.87
28	India	0.2584	4	103.92
29	Mauritania	0.2571	1	122.04
30	Pakistan	0.2504	8	102.51

Table 15.3: All thirty pools, ranked by isolation against weighted-degree centrality, recomputed on the single real, validated 30×30 divergence matrix. Kerala in bold.

Proposition : The Largest Deviation in the Series, Narrowly Exceeding Andhra Pradesh

Kerala's deviation from whole-country India's family-excluded isolation, $+0.0690$, is now the single largest of all sixteen zones in absolute value, narrowly exceeding Andhra Pradesh's own $+0.0677$ and remaining well above the German-speaking Swiss cantons' $+0.0562$, the largest outlier outside the Indian sub-pools. Pearson $r = 0.8675$ ($p = 1.3 \times 10^{-5}$) and Spearman $\rho = 0.5797$ ($p = 0.019$), both a further step down from Punjab's fifteen-zone figures ($r = 0.9083$, $\rho = 0.6385$), continuing the pattern Chapter 13 already reported honestly for Andhra Pradesh: each of the three Indian sub-pools added so far has loosened, not tightened, a correlation built mostly on the twelve non-Indian zones, and Kerala's own deviation — both Indian and Dravidian, unlike Punjab's closer tracking of the Indo-Aryan-majority parent — loosens it furthest of the three.

Delimited zone	Parent	Zone isol.	Parent isol.	Diff.
Central Anatolia (11 prov.)	Turkey	0.3147	0.3006	+0.0141
Cappadocia core (4 prov.)	Turkey	0.3026	0.3006	+0.0020
German-speaking cantons (17)	Switzerland	0.3274	0.2712	+0.0562
Samtskhe-Javakheti	Georgia	0.2734	0.2764	−0.0030
Talysh core (6 rayons)	Azerbaijan	0.2631	0.2629	+0.0002
Northern Ireland	United Kingdom	0.2761	0.2868	−0.0107
Ulster rump	Ireland	0.2904	0.2908	−0.0004
Tripolitania	Libya	0.4194	0.4068	+0.0126
Cyrenaica	Libya	0.4016	0.4068	−0.0052
Fezzan	Libya	0.3965	0.4068	−0.0103
Mauritania (Sahel)	Mauritania	0.2810	0.2737	+0.0073
Mauritania (Sahara)	Mauritania	0.2723	0.2737	−0.0014
Kvemo Kartli	Georgia	0.2818	0.2764	+0.0054
Andhra Pradesh	India	0.3432	0.2755	+0.0677
Punjab	India	0.2950	0.2755	+0.0195
Kerala	**India**	**0.3445**	**0.2755**	+0.0690

Table 15.4: Table 14.4 recomputed on the full thirty-eight-pool universe (thirty-seven plus Kerala). Kerala, sixteenth row, in bold.

15.6 Clustering: The Open Cluster Gains a Tenth Member

UPGMA clustering on the full 30×30 matrix (cophenetic correlation 0.8587, continuing the gradual decline from the twenty-nine-pool figure of 0.8625, the twenty-eight-pool figure of 0.8698, and the twenty-seven-pool figure of 0.8784) still recovers the same five clusters Chapters 5, 8, 13, and 14 report, with Kerala joining cluster (iii): Turkey, Armenia, Georgia, Azerbaijan, India, Pakistan, Kvemo Kartli, Andhra Pradesh, Punjab, and now Kerala — ten members.

Kerala is the fourth sub-national pool, after Kvemo Kartli (Georgia), Andhra Pradesh (India), and Punjab (India), to test the Networks Methodological Volume's "genuinely open" Turkey–Caucasus–India–Pakistan grouping (§11.3 there) by delimitation below the whole-country level, and the fourth to leave the grouping intact rather than pulling either the sub-pool or its parent out of it. Four independent sub-national tests, two of them (Andhra Pradesh, Kerala) now mutually the closest pair of pools in the entire matrix bar their shared parent, continue to agree that delimitation does not dissolve the grouping; none resolves what produces it.

15.7 Scope and Limits of This Extension

Methodological Transparency

Kerala, like Kvemo Kartli, Andhra Pradesh, and Punjab before it, exists at present only within this volume's own 30 × 30 Jensen–Shannon matrix and the family-excluded isolation statistic of §15.5. It is not a vertex in the Methodological Volume's fixed thirty-three-zone centrality graph (Chapter 9 above); it was not part of the persistent-homology sweep of Networks Vol. II, Ch. 14; and it does not appear in the Fisher–Rao/Euclidean distortion result of §4.2. Nor does its addition close the Open Research Agenda item it addresses a third time: one sizeable untested Indian admin1 code remains after Kerala, Punjab, and Andhra Pradesh are all removed from consideration — West Bengal (`28`, 4,309 records) — along with Delhi NCT (`07`, 617 records, smaller than any pool tested so far in this volume) and the still-unverified code `10` (Table 13.5). What this chapter closes is, as before, one further concrete instance of the named agenda item, computed and validated in full, not the item's entire remaining scope.

Chapter 16

A Thirty-First Pool: West Bengal and a New Closeness Record

Methodological Transparency

This chapter addresses a fourth concrete instance of the Open Research Agenda item Chapters 13–15 left explicitly open: the Scope-and-Limits section of Chapter 15 names West Bengal (admin1=28 in IN.txt, 4,309 records) as the last sizeable untested Indian admin1 pool before Delhi NCT and the unverified code 10. It follows the identical protocol the three preceding chapters established: admin1 identification by named-town cross-reference, re-validation of the existing system before extension, then the same two statistics — the Jensen–Shannon isolation/centrality pair and the family-excluded isolation score — and finally UPGMA clustering. No zone definition is invented; no code beyond the admin1 filter already documented in Appendix A is required.

16.1 Admin1 Identification: The Fourth Indian Sub-Pool

Six towns independently known to lie in the Indian state of West Bengal — Kolkata, Howrah, Durgapur, Asansol, Siliguri, and Malda — were located in IN.txt by name, restricted to their highest-population P-class record where a name recurs, exactly as §§13.1, 14.1, and 15.1 did for the three earlier Indian sub-pools. All six, without exception, share admin1=28.

Town	Admin1	Population	Feature code
Kolkata	28	4,631,392	PPLA
Howrah	28	1,027,672	PPL
Durgapur	28	518,872	PPL
Siliguri	28	515,574	PPL
Asansol	28	504,271	PPLA2
Malda	28	170,039	PPLA2

Table 16.1: Admin1 identification for the thirty-first pool: six towns independently known to lie in West Bengal, all sharing admin1=28 in IN.txt, ranked by their highest-population P-class record.

Asansol's own highest-population record in IN.txt carries the diacritic form Āsansol in its asciiname field rather than the plain-ASCII spelling; two zero-population duplicate records under admin1=38 (Bihar) share the plain spelling but are excluded by the same highest-population selection rule already applied to every other town in this and the preceding three Indian chapters. The bigram pipeline itself is unaffected either way: normalize() strips diacritics before extraction, so both spellings would map to the identical normalised string.

16.2 Re-Validation Before Extension

As in Chapters 8, 13, 14, and 15, the thirty-pool system was rebuilt in full before West Bengal was added to any comparison: all thirty isolation and weighted-degree figures of Table 15.3 matched exactly, and Kendall's $\tau = -0.5080$ ($p < 0.0001$) reproduced digit-for-digit. Only with that reproduction confirmed was the thirty-first pool introduced.

16.3 West Bengal's Divergence Profile: The Closest Sub-Pool Yet Recorded

Closest to West Bengal		**Most divergent from West Bengal**	
Pool	D_{JS} (bits)	Pool	D_{JS} (bits)
India (parent)	0.0377	France	0.3595
Punjab	0.0647	Portugal	0.3571
Pakistan	0.1174	Austria	0.3513
Kerala	0.1964	Slovenia	0.3488
Andhra Pradesh	0.2046	Italy	0.3297
Armenia	0.2072	Croatia	0.3190

Table 16.2: West Bengal against all thirty other pools. Its divergence from whole-country India, 0.0377 bits, is the smallest parent/sub-pool divergence yet recorded in this system, smaller than Punjab's own 0.0495 (Table 14.2), the previous record.

West Bengal's own parent, India, is again its closest pool, restoring the pattern Kerala broke in Chapter 15 — but at a distance, 0.0377 bits, under three-quarters of Punjab's own 0.0495, which had itself been the smallest parent/sub-pool divergence recorded anywhere in the system up to that point. West Bengal's second-closest pool is again Punjab, at 0.0647 bits, echoing the Indo-Aryan-cluster pattern Punjab's own §14.3 first reported for its own second-closest neighbour, Pakistan: the two closest non-parent pools to any Indo-Aryan-majority Indian sub-pool tested so far have both been other Indo-Aryan-majority pools (Punjab, Pakistan), while Andhra Pradesh's and Kerala's own second-closest non-parent pools have instead been each other and, further down, Pakistan and Georgia respectively — consistent with, though not proof of, a genuine Indo-Aryan/Dravidian split already visible at the level of raw character bigrams.

Proposition : Four Sub-Pools, Two Distinct Proximity Patterns

The four Indian sub-pools tested across Chapters 13–16 now split cleanly into two pairs by their own closest-neighbour profile. Punjab (0.0495) and West Bengal (0.0377) are both closest to India by a wide margin and second-closest to another Indo-Aryan-majority pool (India, Punjab, respectively for the other); Andhra Pradesh (0.1358) and Kerala (0.1225, the sole case where the parent is not even the closest pool) are both an order of magnitude further from India and closest, first or second, to each other. This is reported strictly as a pattern in the four measured divergences, not as a claim about the underlying Indo-Aryan/Dravidian linguistic classification, which is independently well established on other grounds and is not what this bigram pipeline tests.

16.4 Isolation and Centrality on the Thirty-One-Pool System

Iso. rank	Pool	Isolation (bits)	Degree rank	Weighted degree
1	Tripolitania	0.3758	14	98.91
2	Libya	0.3630	8	106.12
3	Cyrenaica	0.3616	18	97.25
4	Fezzan	0.3566	17	97.28
5	Slovenia	0.3481	31	82.77
6	Croatia	0.3362	28	85.30
7	Austria	0.3332	30	84.76
8	Portugal	0.3323	29	84.95
9	Kerala	0.3238	26	85.43
10	Andhra Pradesh	0.3221	27	85.42
11	Italy	0.3130	24	89.78
12	France	0.3071	22	91.52
13	Turkey	0.3046	25	87.13
14	Ireland	0.2995	21	92.09
15	United Kingdom	0.2948	20	95.32
16	Armenia	0.2946	23	90.28
17	Spain	0.2922	19	95.70
18	Algeria	0.2860	9	105.60
19	Switzerland	0.2778	12	101.25
20	Tunisia	0.2767	7	106.76
21	Kvemo Kartli	0.2758	15	98.14
22	Punjab	0.2743	10	105.56
23	Morocco	0.2689	11	103.84
24	Georgia	0.2685	13	101.04
25	Azerbaijan	0.2679	16	97.31
26	Mauritania (Sahel)	0.2650	2	119.58
27	**West Bengal**	**0.2630**	**5**	**108.78**
28	Mauritania (Sahara)	0.2591	3	116.93
29	Mauritania	0.2576	1	125.15
30	India	0.2510	4	115.32
31	Pakistan	0.2459	6	108.49

Table 16.3: All thirty-one pools, ranked by isolation against weighted-degree centrality, recomputed on the single real, validated 31 × 31 divergence matrix. West Bengal in bold.

Proposition : The Strongest Isolation/Centrality Association Recorded in This System

Kendall's τ between the isolation ranking and the degree-centrality ranking of Table 16.3 is $\tau = -0.5527$ ($p < 0.0001$), the strongest association recorded at any extension in this system so far — stronger than the thirty-pool figure of $\tau = -0.5080$, and stronger than Punjab's own twenty-nine-pool record of $\tau = -0.5172$. West Bengal's own placement is the near-mirror image of Kerala's: isolation rank 27 of 31 (low) against degree rank 5 of 31 (high), among the five most centrally connected pools in the entire system, alongside whole-country Mauritania, its two sub-pools, and Libya. This is the expected signature of a pool with low mean divergence spread broadly across most of the matrix, rather than concentrated closeness to only one or two other pools: West Bengal's own §16.3 shows it closer than 0.21 bits to eight different pools, more than any other Indian sub-pool tested so far.

16.5 Family-Excluded Isolation: A Seventeenth Delimited Zone

India's family for exclusion purposes, extended in Chapter 15 to {India, Andhra Pradesh, Punjab, Kerala}, is extended here to {India, Andhra Pradesh, Punjab, Kerala, West Bengal}: no Indian sub-pool is counted toward any other Indian pool's own isolation average. The full thirty-nine-pool family-excluded universe (thirty-eight plus West Bengal) underlying Table 16.4 was rebuilt from raw GeoNames data and checked against all sixteen published rows before West Bengal was added to any comparison; all sixteen matched within the same small tolerance (≤ 0.0016 bits) Chapters 8, 13, 14, and 15 already report and attribute to one further pool entering every other zone's own averaging set, not to reproduction error.

Delimited zone	Parent	Zone isol.	Parent isol.	Diff.
Central Anatolia (11 prov.)	Turkey	0.3144	0.3002	+0.0142
Cappadocia core (4 prov.)	Turkey	0.3023	0.3002	+0.0021
German-speaking cantons (17)	Switzerland	0.3283	0.2723	+0.0560
Samtskhe-Javakheti	Georgia	0.2715	0.2745	−0.0030
Talysh core (6 rayons)	Azerbaijan	0.2615	0.2613	+0.0001
Northern Ireland	United Kingdom	0.2766	0.2872	−0.0106
Ulster rump	Ireland	0.2907	0.2910	−0.0003
Tripolitania	Libya	0.4161	0.4036	+0.0125
Cyrenaica	Libya	0.3986	0.4036	−0.0050
Fezzan	Libya	0.3934	0.4036	−0.0102
Mauritania (Sahel)	Mauritania	0.2809	0.2736	+0.0073
Mauritania (Sahara)	Mauritania	0.2724	0.2736	−0.0012
Kvemo Kartli	Georgia	0.2799	0.2745	+0.0054
Andhra Pradesh	India	0.3432	0.2755	+0.0677
Punjab	India	0.2950	0.2755	+0.0195
Kerala	India	0.3445	0.2755	+0.0690
West Bengal	**India**	**0.2834**	**0.2755**	+0.0079

Table 16.4: Table 15.4 recomputed on the full thirty-nine-pool universe (thirty-eight plus West Bengal). West Bengal, seventeenth row, in bold.

Proposition : The Smallest Indian Deviation, Closer to Talysh and Samtskhe-Javakheti Than to Its Own Siblings

West Bengal's deviation from whole-country India's family-excluded isolation, +0.0079, is by far the smallest of the four Indian sub-pools — an order of magnitude below Andhra Pradesh's +0.0677 and Kerala's +0.0690, and smaller even than Punjab's own +0.0195. It sits closer in magnitude to the near-zero deviations of the Talysh core (+0.0001) and Samtskhe-Javakheti (−0.0030) than to any other Indian zone. Pearson $r = 0.8653$ ($p = 7.3 \times 10^{-6}$) and Spearman $\rho = 0.6812$ ($p = 0.0026$): the Spearman figure recovers most of the ground lost at Kerala's own addition ($\rho = 0.5797$), since West Bengal's own small, parent-tracking deviation is exactly the kind of data point that tightens rather than loosens a rank correlation already weakened by two larger Dravidian-zone outliers.

16.6 Clustering: The Open Cluster Gains an Eleventh Member

UPGMA clustering on the full 31 × 31 matrix (cophenetic correlation 0.8578, continuing the gradual decline from the thirty-pool figure of 0.8587, the twenty-nine-pool figure of 0.8625, and the twenty-eight-pool figure of 0.8698) still recovers the same five clusters Chapters 5, 8, 13, 14, and 15 report, with West Bengal joining cluster (iii): Turkey, Armenia, Georgia, Azerbaijan, India, Pakistan, Kvemo Kartli, Andhra Pradesh, Punjab, Kerala, and now West Bengal — eleven members.

West Bengal is the fifth sub-national pool, after Kvemo Kartli (Georgia), Andhra Pradesh, Punjab, and Kerala (all India), to test the Networks Methodological Volume's "genuinely open" Turkey–Caucasus–India–Pakistan grouping (§11.3 there) by delimitation below the whole-country level, and the fifth to leave the grouping intact. With four of the cluster's eleven members now Indian sub-pools alone, the cluster's own internal composition is increasingly dominated by within-India delimitation rather than by further whole-country additions — a direct consequence of India being the single largest source pool in `IN.txt` and hence the country whose admin1 codes exhaust least quickly under Chapter 14's own largest-remaining-pool selection rule, not of any property of the cluster itself.

16.7 Scope and Limits of This Extension

Methodological Transparency

West Bengal, like the four delimited zones before it, exists at present only within this volume's own 31 × 31 Jensen–Shannon matrix and the family-excluded isolation statistic of §16.5. It is not a vertex in the Methodological Volume's fixed thirty-three-zone centrality graph (Chapter 9 above); it was not part of the persistent-homology sweep of Networks Vol. II, Ch. 14; and it does not appear in the Fisher–Rao/Euclidean distortion result of §4.2. Its addition does, however, close the sizeable-pool portion of the Open Research Agenda item Chapters 13–15 addressed in turn: after Andhra Pradesh, Punjab, Kerala, and now West Bengal, no untested Indian admin1 code with more than 1,000 records remains in `IN.txt`. What is left is Delhi NCT (`07`, 617 records) and the still-unverified code `10`, whose own top-population entries were already found, in Chapter 13's own Table 13.5, to span more than one state — a genuinely different kind of obstacle (a compound or ambiguous admin1 code) rather than merely a smaller one, and one this chapter does not attempt to resolve.

Chapter 17

A Thirty-Second Pool: Delhi NCT and a Second Non-Parent Closest Neighbour

Methodological Transparency

This chapter addresses a fifth concrete instance of the Open Research Agenda item Chapters 13–16 left explicitly open: the Scope-and-Limits section of Chapter 16 names Delhi NCT (admin1=07 in IN.txt, 617 records) as the last sizeable untested Indian admin1 pool, alongside a still-unverified code, 10. It follows the identical protocol the four preceding chapters established: admin1 identification by named-town cross-reference, re-validation of the existing system before extension, then the same two statistics — the Jensen–Shannon isolation/centrality pair and the family-excluded isolation score — and finally UPGMA clustering. No zone definition is invented; no code beyond the admin1 filter already documented in Appendix A is required. Unlike the four Indian sub-pools before it, Delhi NCT is not a state but a city-state-scale National Capital Territory, and its own record count (617) is smaller than any pool tested so far in this volume — smaller by more than a factor of six than West Bengal's 4,309, the previous smallest.

17.1 Admin1 Identification: The Fifth Indian Sub-Pool

Five towns independently known to lie within the National Capital Territory of Delhi — New Delhi, Delhi, Najafgarh, Narela, and Rohini — were located in IN.txt by name, restricted to their highest-population P-class record where a name recurs, exactly as §§13.1, 14.1, 15.1, and 16.1 did for the four earlier Indian sub-pools. All five, without exception, share admin1=07.

Town	Admin1	Population	Feature code
Delhi	07	11,034,555	PPLA
Najafgarh	07	1,365,000	PPL
Narela	07	800,000	PPL
Rohini	07	860,000	PPLX
New Delhi	07	317,797	PPLC

Table 17.1: Admin1 identification for the thirty-second pool: five towns independently known to lie in the Delhi National Capital Territory, all sharing admin1=07 in IN.txt, ranked by their highest-population P-class record. Population figures are IN.txt's own field, not an external source.

A sixth candidate town, Dwarka, was tested and rejected as a cross-reference for this pool: `IN.txt` carries a Delhi Dwarka record under admin1=`07` with population `0`, and a higher-population Gujarat Dwarka record (38,873, admin1=`09`) that would win the highest-population selection rule used throughout this volume, wrongly assigning the name to the wrong state if applied mechanically. This is the same category of same-name hazard Chapter 15's remark box documented for Asansol's diacritic duplicate and Chapter 13's Table 13.5 documented for admin1 code `10` (§17.7 below); it is handled here, as there, simply by excluding the ambiguous case from the verification set rather than by any special-cased override of the standard rule.

17.2 Re-Validation Before Extension

As in Chapters 8, 13, 14, 15, and 16, the thirty-one-pool system (Chapters 7–8, 13–16) was rebuilt in full from raw GeoNames data and checked against Table 16.3 before Delhi NCT was added to any comparison: all thirty-one isolation and weighted-degree figures matched exactly, and Kendall's $\tau = -0.5527$ ($p = 4.3 \times 10^{-6}$) reproduced digit-for-digit. Only with that reproduction confirmed was the thirty-second pool introduced, and every statistic below is computed fresh over the full 32×32 system.

17.3 Delhi NCT's Divergence Profile: A Second Non-Parent Closest Neighbour

Closest to Delhi NCT		**Most divergent from Delhi NCT**	
Pool	D_{JS} (bits)	Pool	D_{JS} (bits)
West Bengal	0.0688	Portugal	0.3509
India (parent)	0.0858	Austria	0.3357
Punjab	0.0940	Italy	0.3227
Pakistan	0.1253	Slovenia	0.3138
Kvemo Kartli	0.1853	France	0.3123
Armenia	0.1899	Algeria	0.3092

Table 17.2: Delhi NCT against all thirty-one other pools. As with Kerala (Chapter 15), its single closest pool is not its own whole-country parent.

Delhi NCT's closest pool in the full 32×32 matrix is West Bengal, at 0.0688 bits — closer than whole-country India itself, at 0.0858 bits. This is the second case among the five delimited Indian zones now tested in this volume in which a pool's own whole-country parent is not its single closest neighbour: Kerala broke the pattern first (Chapter 15, Andhra Pradesh at 0.1225 bits against India's own 0.1345); Punjab and Andhra Pradesh both had India itself as closest (0.0495 and 0.1358 bits respectively); West Bengal's own closest-pool table (Chapter 16) also listed India as closest, at 0.0377 bits, well clear of any sibling sub-pool. Delhi NCT reverses that: its own margin over India, 0.0170 bits, is smaller than Kerala's 0.0120-bit margin over Andhra Pradesh was, though both are unambiguous within the pipeline's own reproducibility tolerance.

Proposition : An Asymmetric Sibling Relation

West Bengal's own divergence from Delhi NCT, 0.0688 bits, is markedly larger than West Bengal's own divergence from whole-country India, 0.0377 bits (Chapter 16, §16.3): West Bengal is close to India and only moderately close to Delhi, while Delhi is closer to West Bengal than to India. The relation is not symmetric in the sense a naive same-family intuition might expect, because "closest pool to X" and "closest pool to Y" are two different rankings computed from the same distance matrix, not a single mutual-nearest-neighbour relation. Delhi's own small pool size (617 names, against West Bengal's 4,309) is a plausible contributor to why its own aggregate distribution sits, in bigram space, unusually close to a large, populous, single neighbouring pool rather than spread evenly across the wider Indo-Aryan family; this is reported strictly as an empirical distance relation on the character- bigram profiles, not as a claim about which specific dialectal or migratory mechanism produces it, in keeping with this volume's Methodological Transparency convention throughout.

17.4 Isolation and Centrality on the Thirty-Two-Pool System

Proposition : The Strongest Isolation/Centrality Association Yet Recorded

Kendall's τ between the isolation ranking and the degree-centrality ranking of Table 17.3 is $\tau = -0.5806$ ($p = 7.1 \times 10^{-7}$), the strongest association recorded at any extension in this system so far — stronger than the thirty-one-pool figure of $\tau = -0.5527$ (Chapter 16) and stronger than Punjab's own twenty-nine-pool record of $\tau = -0.5172$ (Chapter 14). Delhi NCT's own placement follows West Bengal's pattern rather than Kerala's: isolation rank 30 of 32 (very low) against degree rank 7 of 32 (very high), among the seven most centrally connected pools in the system, immediately behind India and Pakistan themselves. This is the expected signature, given §17.3, of a pool whose mean divergence is pulled down by three very close neighbours (West Bengal, India, Punjab, all under 0.10 bits) rather than one concentrated close pair: Delhi NCT is within 0.13 bits of four different pools, more than any other Indian sub-pool tested so far except West Bengal's own eight.

17.5 Family-Excluded Isolation: An Eighteenth Delimited Zone

India's family for exclusion purposes, extended in Chapter 16 to {India, Andhra Pradesh, Punjab, Kerala, West Bengal}, is extended here to {India, Andhra Pradesh, Punjab, Kerala, West Bengal, Delhi NCT}: no Indian sub-pool is counted toward any other Indian pool's own isolation average. The full forty-pool family-excluded universe (thirty-nine plus Delhi NCT) underlying Table 17.4 was rebuilt from raw GeoNames data — all twenty-six base pools, the seven further Networks Vol. III zones of §7.3, and the five earlier Indian sub-pools — and checked against all seventeen published rows before Delhi NCT was added to any comparison; all seventeen matched within the same small tolerance (≤ 0.0026 bits) Chapters 8, 13, 14, 15, and 16 already report and attribute to one further pool entering every other zone's own averaging set, not to reproduction error.

Iso. rank	Pool	Isolation (bits)	Degree rank	Weighted degree
1	Tripolitania	0.3737	16	101.69
2	Libya	0.3609	10	108.97
3	Cyrenaica	0.3594	19	100.16
4	Fezzan	0.3543	18	100.25
5	Slovenia	0.3470	32	85.51
6	Croatia	0.3351	29	88.16
7	Austria	0.3333	31	87.35
8	Portugal	0.3329	30	87.44
9	Kerala	0.3210	27	88.92
10	Andhra Pradesh	0.3200	28	88.67
11	Italy	0.3133	25	92.46
12	France	0.3073	24	94.28
13	Turkey	0.3038	26	90.16
14	Ireland	0.2979	22	95.42
15	United Kingdom	0.2934	20	98.62
16	Spain	0.2922	21	98.62
17	Armenia	0.2912	23	94.45
18	Algeria	0.2868	11	108.38
19	Switzerland	0.2780	14	104.23
20	Tunisia	0.2770	9	109.72
21	Kvemo Kartli	0.2729	15	102.39
22	Morocco	0.2686	12	107.08
23	Punjab	0.2685	8	112.50
24	Georgia	0.2668	13	104.80
25	Azerbaijan	0.2654	17	101.47
26	Mauritania (Sahel)	0.2652	2	122.69
27	Mauritania (Sahara)	0.2591	4	120.15
28	Mauritania	0.2579	1	128.31
29	West Bengal	0.2568	5	117.20
30	**Delhi NCT**	**0.2513**	**7**	**113.20**
31	India	0.2457	3	122.69
32	Pakistan	0.2420	6	114.20

Table 17.3: All thirty-two pools, ranked by isolation against weighted-degree centrality, recomputed on the single real, validated 32×32 divergence matrix. Delhi NCT in bold.

Proposition : The First Indian Sub-Pool Less Isolated Than Its Own Parent

Delhi NCT's deviation from whole-country India's family-excluded isolation, -0.0093, is the first *negative* deviation recorded among the six Indian sub-pools now tested — every one of Andhra Pradesh ($+0.0677$), Punjab ($+0.0195$), Kerala ($+0.0690$), and West Bengal ($+0.0079$) was, by however small a margin, *more* isolated than whole-country India itself. Delhi NCT is instead marginally *less* isolated from the rest of the forty-pool universe than India is, a pattern so far seen among the Indian zones only in sign, not magnitude: the deviation's absolute size, 0.0093, sits between West Bengal's 0.0079 and Punjab's 0.0195, unremarkable in scale even though its direction is new. Pearson $r = 0.8586$ ($p = 5.1 \times 10^{-6}$) and Spearman $\rho = 0.6552$ ($p = 0.0032$) across the eighteen pairs: both figures move only slightly from the seventeen-pair figures underlying Chapter 16's own Table 16.4 (not separately restated there), consistent with one small-magnitude, sign-reversing addition loosening rather than tightening a rank correlation already carrying two larger Dravidian-zone outliers.

Delimited zone	Parent	Zone isol.	Parent isol.	Diff.
Central Anatolia (11 prov.)	Turkey	0.3138	0.2997	+0.0142
Cappadocia core (4 prov.)	Turkey	0.3014	0.2997	+0.0017
German-speaking cantons (17)	Switzerland	0.3287	0.2726	+0.0560
Samtskhe-Javakheti	Georgia	0.2691	0.2729	−0.0038
Talysh core (6 rayons)	Azerbaijan	0.2590	0.2594	−0.0004
Northern Ireland	United Kingdom	0.2754	0.2863	−0.0109
Ulster rump	Ireland	0.2895	0.2900	−0.0005
Tripolitania	Libya	0.4130	0.4006	+0.0124
Cyrenaica	Libya	0.3956	0.4006	−0.0051
Fezzan	Libya	0.3903	0.4006	−0.0103
Mauritania (Sahel)	Mauritania	0.2806	0.2734	+0.0072
Mauritania (Sahara)	Mauritania	0.2721	0.2734	−0.0013
Kvemo Kartli	Georgia	0.2773	0.2729	+0.0044
Andhra Pradesh	India	0.3432	0.2755	+0.0677
Punjab	India	0.2950	0.2755	+0.0195
Kerala	India	0.3445	0.2755	+0.0690
West Bengal	India	0.2834	0.2755	+0.0079
Delhi NCT	**India**	**0.2662**	**0.2755**	−0.0093

Table 17.4: Table 16.4 recomputed on the full forty-pool universe (thirty-nine plus Delhi NCT). Delhi NCT, eighteenth row, in bold.

17.6 Clustering: The Open Cluster Gains a Twelfth Member

UPGMA clustering on the full 32 × 32 matrix (cophenetic correlation 0.8562, continuing the gradual decline from the thirty-one-pool figure of 0.8578, the thirty-pool figure of 0.8587, and the twenty-nine-pool figure of 0.8625) still recovers the same five clusters Chapters 5, 8, 13, 14, 15, and 16 report, with Delhi NCT joining cluster (iii): Turkey, Armenia, Georgia, Azerbaijan, India, Pakistan, Kvemo Kartli, Andhra Pradesh, Punjab, Kerala, West Bengal, and now Delhi NCT — twelve members.

Delhi NCT is the sixth sub-national pool, after Kvemo Kartli (Georgia), Andhra Pradesh, Punjab, Kerala, and West Bengal (all India), to test the Networks Methodological Volume's "genuinely open" Turkey–Caucasus–India–Pakistan grouping (§11.3 there) by delimitation below the whole-country level, and the sixth to leave the grouping intact. Five of the cluster's twelve members are now Indian sub-pools alone, underscoring §16.4's own remark (Chapter 16) that this is a consequence of India's own admin1-code exhaustion rate in `IN.txt`, not of any property of the cluster itself.

17.7 Resolving Admin1 Code 10: Haryana, Not an Ambiguous Code

Table 13.5 (Chapter 13) left admin1 code `10` explicitly unverified, on the stated grounds that its top-population entries appeared to span more than one state. That impression is addressed here, though the pool itself is not added to any comparison. Eight towns independently known to lie in the Indian state of Haryana — Faridabad, Gurugram, Rohtak, Hisar, Panipat, Karnal, Ambala, and Sirsa — were located in `IN.txt` by the same highest-population protocol used throughout this volume; all eight, without exception, resolve to

admin1=10. Admin1 code 10 is Haryana.

The apparent state-spanning behaviour flagged in Chapter 13 traces to a single same-name collision, not to any genuine ambiguity in Haryana's own delimitation: IN.txt carries a small Haryana village named Gorakhpur (admin1=10, admin2=078) whose population field reads 1,324,570 — the same figure, to the exact digit, as the well-known Uttar Pradesh city of the same name (admin1=36, admin2=188, population 1,324,570 in its own P-class Gorakhpur record). This is almost certainly a data-entry duplication internal to the supplied GeoNames extract rather than two cities genuinely sharing a population, and it is exactly the kind of same-name hazard §17.1's remark box documents for Dwarka and Chapter 15's own remark box documents for Asansol: a name-based scan without the highest-population restriction, or one that happens to select the Haryana Gorakhpur record specifically, would misidentify the pool as spanning Haryana and Uttar Pradesh both. Restricting to towns with no independent cross-state homonym — the eight listed above — removes the hazard entirely. No further computation on the Haryana pool (isolation, centrality, family-excluded isolation, clustering) is attempted in this chapter; identifying admin1=10 is reported as a closed sub-item of Table 13.5's own open list, not as a seventh delimited zone.

17.8 Scope and Limits of This Extension

Methodological Transparency

Delhi NCT, like the five delimited zones before it, exists at present only within this volume's own 32×32 Jensen–Shannon matrix and the family-excluded isolation statistic of §17.5. It is not a vertex in the Methodological Volume's fixed thirty-three-zone centrality graph (Chapter 9 above); it was not part of the persistent-homology sweep of Networks Vol. II, Ch. 14; and it does not appear in the Fisher–Rao/Euclidean distortion result of §4.2. Its addition, together with §17.7's resolution of admin1 code 10 as Haryana, closes the Open Research Agenda item Chapters 13–16 addressed in turn: after Andhra Pradesh, Punjab, Kerala, West Bengal, and now Delhi NCT, no untested Indian admin1 code named in Table 13.5 remains unidentified. Haryana itself (admin1=10, comparable in record count to Kerala) is now identified but not yet tested on any of this volume's own statistics — a natural next instance, should the agenda item be reopened, but a genuinely new one, not a residue of the item as stated in Chapter 13.

Chapter 18

A Thirty-Third Pool: Haryana and the Closest Pair Yet Recorded

Methodological Transparency

This chapter was not part of this volume's original closing sequence. Chapter 20 below records, in its own words, a deliberate decision to stop the Indian sub-pool extension at six tested instances and one identified, untested one — Haryana, admin1=10 in IN.txt, resolved in §17.7 but explicitly not run through this volume's own pipelines, on the stated ground that a half-finished computation, patched in under session-time pressure, would have cost more in methodological consistency than one further data point was worth. That constraint does not apply in the present session, in which the raw GeoNames extract and the full working time needed to rebuild the thirty-two-pool system from source, add Haryana, and validate every existing published figure before trusting the extension were all available together. This chapter is the result. It changes nothing about the reasoning Chapter 20 gives for why six-plus-one was a defensible place to stop in that earlier session; it reports what running the seventh instance the closing chapter left on the table actually shows, now that running it was possible. Chapter 20 itself is left as it was written, with a short remark added at the relevant point rather than rewritten, following this volume's own stated preference for flagging a tension explicitly over smoothing it away.

18.1 Admin1 Identification: Already Resolved, Not Repeated

Haryana's admin1 code was identified in §17.7 (Chapter 17) by the same eight-town, highest-population cross-reference protocol used for every other Indian sub-pool in this volume — Faridabad, Gurugram, Rohtak, Hisar, Panipat, Karnal, Ambala, and Sirsa, all resolving to admin1=10 in IN.txt, with the apparent cross-state ambiguity that originally flagged the code in Chapter 13 traced there to a single duplicated-population Gorakhpur record, not a genuine multi-state spread. Nothing in that identification is redone here; this chapter begins from admin1=10 as an already-closed sub-item and proceeds directly to testing the pool it denotes.

18.2 Re-Validation Before Extension

As in Chapters 8 and 13–17, the thirty-two-pool system (Chapters 7–8 and 13–17) was rebuilt in full from the raw GeoNames files before Haryana was added to any comparison. All thirty-two isolation and weighted-degree figures of Table 17.3 matched exactly, and Kendall's $\tau = -0.5806$ ($p = 7.1 \times 10^{-7}$) reproduced digit-for-digit. Only with that reproduction confirmed was the thirty-third pool introduced, and every statistic below is computed fresh over the full 33×33 system, not patched onto the published 32×32 one.

18.3 Haryana's Divergence Profile: The Closest Pair Recorded So Far

Haryana's pool (admin1=10, feature classes P/T, normalised length ≥ 3) contains 7,245 names, comparable in size to Punjab's 11,508 and larger than Kerala's 7,153 or Delhi NCT's 617. Haryana's own row of the 33×33 matrix, sorted ascending, gives:

Pool	D_{JS} to Haryana (bits)
Punjab	0.0254
India	0.0428
West Bengal	0.0508
Delhi NCT	0.0747
Pakistan	0.0940
Armenia	0.2210
Georgia	0.2228
Kerala	0.2255

Table 18.1: Haryana's eight closest pools in the 33×33 matrix.

Proposition : A New Minimum, Not Merely a Third Non-Parent Case

Haryana's closest pool is Punjab, at $D_{JS} = 0.0254$ bits, not whole-country India (0.0428) — a third instance in this system of a delimited zone whose nearest neighbour is not its own parent, after Kerala (Chapter 15, against Andhra Pradesh) and Delhi NCT (Chapter 17, against West Bengal). What is new here is not the non-parent pattern itself but the magnitude: 0.0254 bits is smaller than any pairwise divergence reported anywhere else in this volume, including the twenty-six-pool matrix's own most similar pair, Algeria–Tunisia at 0.0347 (Table 5.1), and Tripolitania–Fezzan at 0.0355. Punjab–Haryana is, on the evidence assembled in this volume's own tables, the closest pair of pools recorded at any stage of either series. This is a plausible rather than a surprising result — Punjab and Haryana are contiguous present-day Indian states carved from a shared pre-Partition Punjab region, sharing substantial toponymic vocabulary by administrative and settlement history rather than by any deeper substrate claim — but it is reported as a new extreme because no earlier chapter's own extremes table had reason to anticipate a pair this close.

Iso. rank	Pool	Isolation (bits)	Degree rank	Weighted degree
1	Tripolitania	0.3711	18	104.63
2	Libya	0.3584	11	112.02
3	Cyrenaica	0.3569	20	103.19
4	Fezzan	0.3521	19	103.26
5	Slovenia	0.3477	33	87.90
6	Croatia	0.3353	30	90.70
7	Portugal	0.3346	32	89.73
8	Austria	0.3342	31	89.76
9	Kerala	0.3180	28	92.54
10	Andhra Pradesh	0.3177	29	92.06
11	Italy	0.3149	26	94.87
12	France	0.3096	25	96.60
13	Turkey	0.3041	27	92.92
14	Ireland	0.2984	23	98.16
15	United Kingdom	0.2942	21	101.33
16	Spain	0.2937	22	101.20
17	Armenia	0.2890	24	98.14
18	Algeria	0.2879	12	111.07
19	Switzerland	0.2796	15	106.87
20	Tunisia	0.2780	10	112.52
21	Kvemo Kartli	0.2715	16	105.98
22	Morocco	0.2688	13	110.13
23	Mauritania (Sahel)	0.2660	6	125.62
24	Georgia	0.2654	14	108.46
25	Azerbaijan	0.2642	17	105.06
26	**Haryana**	**0.2627**	**3**	**127.53**
27	Punjab	0.2609	5	125.77
28	Mauritania (Sahara)	0.2605	7	122.98
29	Mauritania	0.2589	2	131.26
30	West Bengal	0.2503	4	127.11
31	Delhi NCT	0.2458	8	121.22
32	India	0.2393	1	133.46
33	Pakistan	0.2374	9	121.14

Table 18.2: All thirty-three pools, ranked by isolation against weighted-degree centrality, recomputed on the single real, validated 33×33 divergence matrix. Haryana in bold.

18.4 Isolation and Centrality on the Thirty-Three-Pool System

The isolation and weighted-degree *values* of Table 18.2 were recomputed directly from the 33×33 matrix and matched, on first computation, in every cell. The Degree rank column, however, did not: a first pass of this table carried several ranks over from stale intermediate ordering rather than resorting the weighted-degree column freshly once Haryana's row was inserted, producing errors as large as Pakistan's (originally listed at rank 32, an implausible value alongside its own weighted degree of 121.14, actually rank 9) and smaller one-to-two-place errors elsewhere (e.g. Mauritania (Sahel), West Bengal, Azerbaijan). The table above reflects the corrected ranking, resorted directly from the weighted-degree column already printed beside it; no isolation value, weighted-degree value, or any other statistic in this chapter was affected, since the error was confined to this one derived column.

Proposition : A New Record, by a Small Margin

Kendall's τ between the isolation ranking and the degree-centrality ranking of Table 18.2 is $\tau = -0.5909$ ($p = 2.6 \times 10^{-7}$), the strongest association recorded at any extension in this system so far — marginally stronger than the thirty-two-pool figure of $\tau = -0.5806$ (Chapter 17). The increment, -0.0103, is smaller than four of the preceding six steps in Table 21.1, consistent with the pattern that chapter's own propositionbox already noted: an uncontrolled sequence of additions that mostly, but not uniformly, strengthens the association. Haryana's own placement — isolation rank 26 of 33 (moderately low), degree rank 3 of 33 (very high, exceeded only by India and Mauritania) — follows the same broadly-spread-closeness mechanism West Bengal and Delhi NCT already exhibit rather than the concentrated-divergence mechanism of Andhra Pradesh and Kerala: Haryana sits within 0.075 bits of four different pools (Table 18.1), the same signature Chapter 17 identified for Delhi NCT.

18.5 Family-Excluded Isolation: A Nineteenth Delimited Zone

India's family for exclusion purposes, extended in Chapter 17 to {India, Andhra Pradesh, Punjab, Kerala, West Bengal, Delhi NCT}, is extended here to {India, Andhra Pradesh, Punjab, Kerala, West Bengal, Delhi NCT, Haryana}. The full forty-one-pool family-excluded universe (forty plus Haryana) was rebuilt from raw GeoNames data — the twenty-six base pools, the seven further Networks Vol. III zones of §7.3, and the six earlier Indian sub-pools — and checked against all eighteen published rows of Table 17.4 before Haryana was added to any comparison; all eighteen matched exactly, with zero drift for the six existing Indian-family rows themselves, since Haryana joins their own family and is therefore excluded from their averaging set rather than added to it — the same zero-drift pattern this volume's own family-exclusion definition predicts whenever a new pool joins an existing family, as distinct from the small, non-zero drift every non-Indian row shows when a new pool outside its family enters its own averaging set.

Delimited zone	Parent	Zone isol.	Parent isol.	Diff.
Central Anatolia (11 prov.)	Turkey	0.3143	0.3000	+0.0143
Cappadocia core (4 prov.)	Turkey	0.3018	0.3000	+0.0018
German-speaking cantons (17)	Switzerland	0.3298	0.2742	+0.0556
Samtskhe-Javakheti	Georgia	0.2679	0.2716	−0.0036
Talysh core (6 rayons)	Azerbaijan	0.2581	0.2586	−0.0005
Northern Ireland	United Kingdom	0.2765	0.2872	−0.0107
Ulster rump	Ireland	0.2902	0.2906	−0.0004
Tripolitania	Libya	0.4096	0.3972	+0.0124
Cyrenaica	Libya	0.3924	0.3972	−0.0049
Fezzan	Libya	0.3873	0.3972	−0.0099
Mauritania (Sahel)	Mauritania	0.2809	0.2738	+0.0071
Mauritania (Sahara)	Mauritania	0.2729	0.2738	−0.0009
Kvemo Kartli	Georgia	0.2760	0.2716	+0.0044
Andhra Pradesh	India	0.3432	0.2755	+0.0677
Punjab	India	0.2950	0.2755	+0.0195
Kerala	India	0.3445	0.2755	+0.0690
West Bengal	India	0.2834	0.2755	+0.0079
Delhi NCT	India	0.2662	0.2755	−0.0093
Haryana	**India**	**0.2982**	**0.2755**	+0.0227

Table 18.3: Table 17.4 recomputed on the full forty-one-pool universe (forty plus Haryana). Haryana, nineteenth row, in bold.

Proposition : A Mid-Range Deviation, Third-Largest of Seven

Haryana's deviation from whole-country India's family-excluded isolation, +0.0227, is the third-largest of the seven Indian instances now tested — smaller than Kerala's +0.0690 and Andhra Pradesh's +0.0677 by a wide margin, but larger than Punjab's own +0.0195, West Bengal's +0.0079, and Delhi NCT's −0.0093. Pearson $r = 0.8529$ ($p = 3.5 \times 10^{-6}$) and Spearman $\rho = 0.7276$ ($p = 0.0004$) across the nineteen pairs: both figures move only slightly from the eighteen-pair figures underlying Chapter 17's own Table 17.4, and the Spearman figure in particular strengthens rather than loosens, since Haryana's mid-range deviation sits closer to the eighteen-pair median than either of the two Dravidian-zone outliers that have driven this statistic's own leverage since Chapter 13.

18.6 Clustering: The Open Cluster Gains a Thirteenth Member

UPGMA clustering on the full 33 × 33 matrix (cophenetic correlation 0.8577) recovers the same five clusters Chapters 5, 8, 13–17 report, with Haryana joining cluster (iii): Turkey, Armenia, Georgia, Azerbaijan, India, Pakistan, Kvemo Kartli, Andhra Pradesh, Punjab, Kerala, West Bengal, Delhi NCT, and now Haryana — thirteen members. The cophenetic figure itself breaks, rather than extends, the gradual decline Chapter 17 traced across five successive additions (0.8625 → 0.8587 → 0.8578 → 0.8562): the thirty-three-pool figure, 0.8577, is a small increase over the thirty-two-pool figure of 0.8562, the first such reversal recorded in this sequence. Consistent with Chapter 20's own observation about τ's single non-monotonic step (Kerala, §21.1), one reversal in an otherwise one-directional run of six steps is not treated here as evidence against the general pattern, only as a second data point showing the pattern was never guaranteed by the method.

Haryana is the seventh sub-national pool, after Kvemo Kartli (Georgia), Andhra Pradesh, Punjab, Kerala, West Bengal, and Delhi NCT (all India), to test the Networks Methodological Volume's "genuinely open" Turkey–Caucasus–India–Pakistan grouping (§11.3 there) by delimitation below the whole-country level, and the seventh to leave the grouping intact. Six of the cluster's thirteen members are now Indian sub-pools alone, a proportion this volume attributes, as Chapter 17 already did for five, to India's own admin1-code exhaustion rate in `IN.txt` rather than to any property of the cluster itself.

18.7 Scope and Limits of This Extension

Methodological Transparency

Haryana, like the six sub-pools before it, exists at present only within this volume's own 33×33 Jensen–Shannon matrix and the family-excluded isolation statistic of §18.5. It is not a vertex in the Methodological Volume's fixed thirty-three-zone centrality graph of Chapter 9 above, which is a different, fixed thirty-three-zone system defined by the Entropy series' own original delimitation and unrelated to this chapter's admin-verified pool of the same cardinality; the coincidence of both systems reaching size thirty-three is not a convergence of any substantive kind and is noted here only to avoid the two being conflated. With Haryana's addition, every admin1 code Table 13.5 (Chapter 13) originally flagged as untested or unidentified is now both identified and tested; no further named residue of that table's own list remains at any stage. India retains, at admin1 resolution, roughly two dozen further states and union territories this volume has never touched, and nothing in the method used for Andhra Pradesh through Haryana would stop it from being applied to any of them — the same extensible-sequence observation Chapter 20 made about the six-instance run applies undiminished to the seven-instance one. This chapter does not reopen that question; it closes the one specific, previously-identified-but-untested instance Chapter 20 left on the table, and stops there.

Chapter 19

Open Research Agenda

The following items are compiled, without addition, from the closing chapters of the ten source volumes. Each is attributed to its origin.

- **Fine etymological validation.** A Celtic-annotated sub-corpus of Iberian toponyms with established Celtic etymologies (Villar 1995; de Hoz 2011), to test the Galician–Catalan proximity at finer resolution than the whole-zone level. *(Entropy, Vol. II)*
- **Diachronic substrate erosion.** Measuring assimilation rate directly from dated historical corpora (e.g. thirteenth-century Alfonsine cadastral records) against the modern GeoNames Arabic–Castilian divergence. *(Entropy, Vol. II)*
- **Phoneme-level / IPA representation**, replacing ASCII normalisation, expected to strengthen the Dravidian retroflex signal and Kartvelian/Armenian distinctiveness. *(Entropy, Vols. V, VI)*
- **Targeted stratum analysis** restricted to the oldest epigraphic or scriptural toponymic layer of the Indus basin, as a sharper test of the Dravidian-substrate hypothesis for the Harappan civilisation than the aggregate modern corpus allows. *(Entropy, Vol. V)*
- **Explicit spatial modelling** of substrate-zone boundaries (Gaussian-process boundaries, spatial mixture models), replacing the current discrete administrative-boundary proxy. *(Entropy, Vol. VI)*
- **Extension to the Horn of Africa** (Cushitic, Coptic, Ethiosemitic substrates), to test the generality of the Arabic-substrate isolation pattern observed in Libya against a different Afroasiatic branch. *(Entropy, Vol. VII, explicitly proposed as one of two possible directions for a "Volume VIII")*
- **Cross-corpus KL re-fitting** of all zones over a single shared vocabulary, to replace the entropy-derived proxy inter-regional distances with a genuine 33×33 (or 26×26) directly measured KL matrix. *(Networks, Methodological Volume, §17.2)*
- **GeoNames admin1 cross-reference table** (Swiss cantons, Croatian županije) to map Networks Vol. II's anonymous admin1 strata back to named regions. *(Networks, Vol. II, §18, Methodological Transparency box)*
- **Resolving the Caucasus–Turkey–India–Pakistan cluster**, whose genealogical or areal basis remains, in the source volume's own words, "genuinely open." *(Networks, Methodological Volume, §11.3)*
- **Admin2-level refinement** of zones so far delimited only at admin1 resolution: Trentino-Alto Adige, historic Occitania, Croatian Istria. *(Networks, Vol. III, §29.5)*

☒ **Resampling checks** on small-n pools whose distinctiveness had not yet been separated from sampling noise: Islamabad, Nouakchott. This is no longer open: Chapter 10 above runs the Networks Vol. III §12.1 $B = 12$ protocol on Islamabad directly (an honest negative result — the published figure does not exceed the resampling noise floor) and

explains why the identical protocol cannot run on Nouakchott's 8-entry pool, reporting a population-level JSD comparison and a leave-one-out jackknife as the closest honest substitute, with an explicit methodological caution that both figures are dominated by Laplace-prior smoothing at this sample size rather than by genuine toponymic signal. *(Networks, Vol. III, §29.5; Synthesis, Ch. 10, this volume)*

☒ **Extension to untested admin1 units** of already-delimited countries, of which India's remaining unstudied states are the single largest instance. Now resolved in full. Chapter 13 adds Andhra Pradesh (admin1=02 in IN.txt), the single largest untested Indian admin1 pool by record count, tested on the full isolation/centrality and family-excluded-isolation pipelines and found to be the system's new lowest-centrality pool and the largest single deviation yet recorded on the family-excluded statistic. Chapter 14 adds Punjab (admin1=23), the second-largest, found instead to be the system's closest cross-border pair with whole-country Pakistan (ninth-smallest divergence in the entire matrix, parent/child pairs excluded) and to strengthen, rather than merely reproduce, the isolation/degree association. Chapter 15 adds Kerala (admin1=13), the third-largest, found to be the first delimited zone in the entire system whose closest pool is not its own whole-country parent — Andhra Pradesh, its Dravidian-substrate sibling, is marginally closer — and to record the largest single deviation yet on the family-excluded statistic, narrowly exceeding Andhra Pradesh's own. Chapter 16 adds West Bengal (admin1=28), the fourth-largest and smallest deviation of the four, found to be the system's most centrally connected Indian sub-pool and to close the sizeable-pool (>1,000 records) portion of the item. Chapter 17 adds the last two named residues directly: Delhi NCT (admin1=07, 617 records), found to be a second delimited zone whose closest pool is not its own parent — West Bengal, not whole-country India — and the first Indian sub-pool with a *negative* family-excluded-isolation deviation from its parent; and the previously unverified code 10, resolved by eight-town cross-reference to Haryana, its earlier ambiguous appearance traced to a single Gorakhpur name-collision rather than any genuine multi-state spread. Chapter 18, added in a later session once the working time to do so without a half-finished computation became available, tests Haryana itself on the full isolation/centrality and family-excluded pipelines, finding it the closest pair of pools recorded anywhere in this volume (Punjab–Haryana, $D_{JS} = 0.0254$ bits) and a third non-parent-closest-neighbour case. Seven concrete Indian sub-national instances are now computed and validated in full, every one tested on every statistic, and the item's own named residue list (Table 13.5) is exhausted with no untested residue remaining. *(Networks, Vol. III, §29.5; Synthesis, Chs. 13–17 and 18, this volume)*

☒ **Extending the $B = 12$ resampling protocol to every cross-pool ranking claim** in *Networks*, Vol. III, of which Sindh (§16.3), Corsica (§17.3), and Sardegna (§19.3) were named directly as candidates. Partly resolved: Chapter 11 above runs the identical protocol on all three, finding Sindh and Corsica the source of the resampled matrix's overall maximum in thirteen of thirteen independent draws each, and Sardegna in twelve of thirteen (the sole exception a 0.0002-radian near-tie). The wider "systematic programme" the roadmap item describes — every ranking claim in the volume, not only the three named candidates — remains open beyond these three. *(Networks, Vol. III, §20.3; Synthesis, Ch. 11, this volume)*

☒ **The one quantitative item of Chapter 7 that remained open at the time of that chapter's drafting**: recomputing the Methodological Volume's original proxy-weighted 33-zone graph (Ch. 9 there) with the Entropy series' thirty-three zones defined exactly as that series originally delimited them, over a single shared vocabulary. This is no longer

open: Chapter 9 above compiles the admin1/admin2 code sets for all thirty-three zones (twenty-one of them documented here for the first time in this volume, traced to their own Entropy-series source chapters) and recomputes the full 33×33 divergence matrix and its centrality structure directly. *(Synthesis, Ch. 9, this volume)*

Methodological Transparency

All three items of the original three-item list are now resolved. The first two siblings — a direct Entropy/Networks rank-correlation on a shared statistic, and a delimited-zone/whole-country correlation — were computed directly from raw GeoNames data once that data became available in this session, and are reported with real figures in Chapter 7, §§7.2–7.3 above. The third and last, the full thirty-three-zone graph recomputed on directly measured divergences rather than the original proxy formula, is reported with real figures in Chapter 9 above. None of the three is merely proposed any longer.

The item above concerning "resolving the Caucasus–Turkey–India–Pakistan cluster" remains open in its original genealogical sense, but no longer untouched: Chapter 8 tests it against one additional, independently obtained sub-national pool (Kvemo Kartli) and finds the cluster robust to that one test, without resolving its genealogical or areal basis; Chapter 9 adds, incidentally, that Anatolian and Caucasian-TR — two of the four zones in that cluster — are in fact the single closest pair in the entire thirty-three-zone graph once directly measured, which sharpens without resolving the question. This is reported there, not here, since it is a genuine partial result and not, like the three items above, a completed answer to the item as originally stated.

A fifth item, listed separately above, is also now resolved: the Islamabad/Nouakchott resampling checks. Unlike the four items just discussed, this one resolves asymmetrically rather than completely — Islamabad's half returns a clean, checkable answer (a negative result: its published figure does not exceed the resampling noise floor); Nouakchott's half resolves instead into a precise account of why the named protocol cannot be run at all on an 8-entry pool, together with the closest honest substitute the same 8 entries support. Chapter 10 above reports both halves in full, including the specific methodological reason — Laplace-prior dominance at this sample size — that the substitute diagnostics cannot be read as either confirming or disconfirming Nouakchott's distinctiveness.

A sixth item is also now resolved, fully rather than partially: extension to untested Indian admin1 units. Chapters 13–16 added Andhra Pradesh, Punjab, Kerala, and West Bengal in turn, each tested on the full isolation/centrality and family-excluded pipelines; Chapter 17 closed the two remaining named residues from Chapter 13's own Table 13.5 — Delhi NCT, tested on every statistic and found to be a second non-parent-closest-neighbour case and the first negative family-excluded deviation among the Indian sub-pools tested to that point, and admin1 code `10`, identified as Haryana by an eight-town cross-reference that also traces the code's earlier apparent ambiguity to a single duplicated-population record for a minor same-named village rather than any real multi-state spread. Haryana's own admin1 pool, identified but not yet tested at the time Chapter 17 was written, is tested in full in Chapter 18 below, added once a subsequent session made completing the computation without time-pressure shortcuts possible: it is the closest pair of pools recorded anywhere in this volume (against Punjab, $D_{JS} = 0.0254$ bits) and a third non-parent-closest-neighbour case. The item is now exhausted both as originally scoped and as a source of further already-identified sub-pools; what remains, as Chapter 18's own closing section states, is the much larger and explicitly out-of-scope space of India's remaining untested admin1 units, not a residue of this specific item.

Chapter 20

Closing Reflections: What Kind of Closure a Synthesis Volume Can Offer

The preceding chapters extended this volume five times over — Chapters 8, 13, 14, 15, 16, and 17 each added one further admin-verified pool to the same isolation/centrality and family-excluded-isolation pipelines, following a protocol fixed at Chapter 7 and never altered since. This chapter does not add a sixth. It asks instead what that six-chapter run has actually established, what it has not, and why the run stops here rather than at some later, equally defensible point.

This chapter is preserved as it was written, in the session that closed the Indian sub-pool sequence at six tested instances and one identified, untested one. A later session, with the working time to rebuild the full system and validate it before extending it available in a single sitting rather than under a hard session-time constraint, ran the seventh instance — Haryana — this chapter's own §20.2 declined to run. That result is reported in Chapter 18, inserted immediately before the Open Research Agenda chapter above. It does not retroactively change the reasoning below, which was sound given the constraint it was written under; it does mean the phrase "six-chapter run" in this chapter refers to the state of the volume at the time this chapter was drafted, not to the volume's present, seven-chapter state. Where a specific numbered claim below (chapter counts, the closing remarkbox on Haryana specifically) would otherwise read as still current, it is left unedited rather than silently updated, and this note is offered as the single place a reader needs to reconcile it against Chapter 18.

20.1 The Difference Between an Open Item and an Extensible One

Table 13.5's own residue list — the untested Indian admin1 codes named in Chapter 13 — is now exhausted in the literal sense: every code that table flagged has been either tested (Andhra Pradesh, Punjab, Kerala, West Bengal, Delhi NCT) or identified and explicitly held back from testing (Haryana, §17.7). But exhausting a named list is not the same thing as exhausting the underlying possibility. India has, at admin1 resolution, roughly two dozen further states and union territories this volume has never touched; nothing in the method used for Andhra Pradesh through Delhi NCT would stop it from being applied to any of them, or, for that matter, to any admin1 subdivision of any of the other thirty-two pools. The same is true one level down: Chapter 12's admin2 refinement of Trentino-Alto Adige and Occitania could in principle be attempted on every admin1 zone in the matrix that has admin2 substructure in its own GeoNames extract.

This is the distinction this closing chapter wants to name explicitly, because the six-chapter run risked obscuring it: an *open item* is one where a specific, boundable question has a specific, boundable answer not yet obtained — the Caucasus–Turkey–India–Pakistan cluster's genealogical basis (§19, still open after three independent sub-national tests) is open in this sense, because no further test of the same kind would resolve it; what would resolve it is a different kind of evidence entirely, external to this corpus. An *extensible item*,

by contrast, is one where the method itself defines an unbounded sequence of further instances, each individually cheap to add and individually informative in some small way, with no internal stopping rule. The Indian sub-pool sequence was extensible in exactly this sense. Six instances were run; a seventh (Haryana) was already staged, half-computed, in the working session that produced this volume, when that session's own limits — not any property of the data or the question — ended the run. That the run stopped at a session boundary rather than at a principled boundary is itself the finding this chapter records.

20.2 What the Extensible Run Was Worth, and Where Its Returns Fell

The six Indian instances were not worthless; three of them changed the qualitative picture in ways the first two did not anticipate. Andhra Pradesh (Ch. 13) set a record for isolation/centrality mismatch and for family-excluded deviation that nothing since has matched. Punjab (Ch. 14) found the closest cross-border pair in the entire thirty-three-plus-pool system and the single strongest isolation/degree association recorded in this volume, $\tau = -0.5172$. Kerala (Ch. 15) broke a rule — closest pool equals whole-country parent — that every one of the twenty-nine pools tested before it had obeyed without exception. Delhi NCT (Ch. 17) repeated that break independently and added the first negative family-excluded deviation among the six. West Bengal (Ch. 16), by contrast, mostly reproduced the pattern already established: a real result, but the least informative of the five, since it confirmed rather than complicated the picture. This unevenness is itself instructive: an extensible sequence does not yield diminishing returns uniformly or predictably, which is exactly why it is tempting to keep extending it in search of the next Kerala. It is also why a rule for stopping cannot be "stop when the returns are small," since that is not knowable in advance; it has to be a rule about scope, decided beforehand, and honoured even when the next instance might have been another Kerala.

Nothing in this volume's own statistics changes if a seventh Indian instance, Haryana, is never run. Haryana is identified (admin1=10, §17.7) to the same standard of evidence as the six pools already tested, and a reader who wants the isolation, centrality, and family-excluded-isolation figures for it can reproduce them directly from Appendix A's unchanged pipeline with the single filter `{'10'}`, exactly as Chapters 13–17 did for their own pools. This volume reports that the computation is reproducible and well-defined, not that it was completed, and treats the difference as material: a half-finished extension, patched in under time pressure, would have degraded this volume's own standard of "every figure traced to a chapter that ran it end to end" for the sake of one more data point the closing item did not strictly require.

In the session that produced this chapter, the constraint above was real: completing the seventh instance properly, with full re-validation of the existing thirty-two-pool system first, was not achievable without either rushing the re-validation step or risking an unfinished chapter. In a later session, working from the same raw GeoNames files with the time to do the re-validation properly, the computation described above as reproducible-but-not-completed was in fact completed. Chapter 18 reports it in full: Haryana's own isolation (rank 26 of 33), its weighted-degree centrality (rank 3 of 33), its family-excluded deviation from India ($+0.0227$, third-largest of seven), and its closest pool — Punjab, at $D_{JS} = 0.0254$ bits, the closest pair of pools recorded anywhere in either series to date. None of this overturns the reasoning above; the decision to stop at six-plus-one was the right decision given the constraint it was made under, and the constraint, not the decision, is what changed.

20.3 A Scope Decision, Stated Rather Than Left Implicit

This volume's own stated purpose, set out in the Preface, is to close two prior series' balance sheets, not to become a third series in its own right. Chapters 7 through 9 did that closing work directly, on items the prior volumes had themselves named as open. Chapters 8 and 13–17 went further than that purpose strictly required, because the Indian sub-pool item, once opened, kept producing genuinely new patterns rather than merely confirming old ones — which is a good reason to have run six instances, and an equally good reason not to treat the sequence's non-termination as evidence that a seventh, eighth, or twentieth instance remains owed. The decision recorded here is to close this volume's own extensible sequence at six tested Indian instances and one identified, untested one, and to redirect any further effort toward consolidating what these seventeen chapters already establish rather than toward a thirty-third, thirty-fourth, or thirty-fifth pool. Nothing about the method precludes a future volume from resuming the sequence; nothing in the results obtained so far makes that resumption urgent.

Chapter 21

Trajectory, Not Instance: Eight Figures, Assembled Rather Than Extended

Methodological Transparency

This chapter, as originally drafted, assembled seven values of one statistic and five of a second, out of Chapter 8 and Chapters 13–17. Chapter 18, added in a later session, supplied an eighth value of the first statistic and a sixth of the second; both are folded into the tables below, and §21.3's original predictive discussion, written before that eighth value existed, is kept alongside a direct comparison against what the eighth value actually turned out to be. Every other number below was already published, once, in Chapter 8, Chapters 13–17, or Chapter 18, each time as a single figure reported alongside the pool that produced it. What this chapter does is different in kind from what those chapters did: it takes eight already-validated values of one statistic, and six already-validated values of a second, out of the eight separate chapters that reported them one at a time, and reads them as the single ordered sequence each of those chapters' own prose gestured toward — "markedly stronger," "a small step back," "the strongest yet recorded" — but never assembled into one table. No GeoNames file is opened in this chapter; no pipeline is run; no figure here can be found by re-executing anything. Every entry in Tables 21.1 and 21.3 below carries a citation to the chapter and table it was first reported in, and none is restated to a different number of decimal places than its source.

21.1 A Statistic Reported Eight Times, Read Once

Table 7.2 through Table 18.2 each report Kendall's τ between isolation rank and weighted-degree-centrality rank on the full pool system available at that chapter, immediately after one further pool is added. Eight such figures now exist, spanning Chapter 7's original twenty-six-pool base through Chapter 18's thirty-three-pool system. No source chapter tabulates more than two of these figures side by side — each compares its own result only to its immediate predecessor, which is the correct scope for a chapter arguing that one specific addition behaves consistently with the last one, but leaves the eight-point series itself unassembled anywhere in the volume.

Pools (N)	Addition	τ	p	Source
26	(base system)	-0.4092	0.0030	Ch. 7, Table 7.2
27	Kvemo Kartli	-0.4701	0.0004	Ch. 8, Table 8.3
28	Andhra Pradesh	-0.4709	0.0003	Ch. 13, Table 13.3
29	Punjab	-0.5172	< 0.0001	Ch. 14, Table 14.3
30	Kerala	-0.5080	< 0.0001	Ch. 15, Table 15.3
31	West Bengal	-0.5527	< 0.0001	Ch. 16, Table 16.3
32	Delhi NCT	-0.5806	7.1×10^{-7}	Ch. 17, Table 17.3
33	Haryana	-0.5909	2.6×10^{-7}	Ch. 18, Table 18.2

Table 21.1: The isolation/degree-centrality Kendall's τ, as reported at each of the eight points in this volume where one further pool was added to the system. Every figure is copied unchanged from its source table; none is recomputed here.

Proposition : One Reversal in Seven Steps, and What It Does and Does Not License

Six of the seven steps in Table 21.1 strengthen the association; one, Kerala's addition at $N = 30$, weakens it slightly against Punjab's immediately preceding figure (-0.5080 against -0.5172) before West Bengal and Delhi NCT resume the strengthening trend and surpass Punjab's own record twice over. This is not evidence of a noisy or unreliable statistic: τ is being recomputed on a different $N \times N$ matrix at every step, over a base system that was never designed as a controlled sequence — the six Indian additions entered in the order Table 13.5 (Chapter 13) happened to name them, not in an order chosen to produce a monotone trend, and a single reversal in seven uncontrolled steps is what that kind of sequence should be expected to show from time to time rather than evidence requiring its own separate explanation.

Pool	Isolation rank/N	Degree rank/N
Andhra Pradesh	0.214	1.000
Punjab	0.586	0.448
Kerala	0.300	0.967
West Bengal	0.871	0.161
Delhi NCT	0.938	0.219
Haryana	0.788	0.091

Table 21.2: Isolation and weighted-degree ranks of the six Indian sub-pools, each normalized by dividing by its own chapter's system size N. A value near 0 for isolation rank/N denotes high isolation; a value near 0 for degree rank/N denotes low centrality. Source ranks are quoted verbatim from Chapters 13–17 and 18.

Read this way, the six tested pools separate into two groups Chapter 19's own closing chapter named descriptively but did not tabulate. Andhra Pradesh and Kerala sit at or near the isolation-rank floor and the degree-rank ceiling simultaneously (0.214/1.000 and 0.300/0.967): both are moderately-to-highly isolated pools whose closeness to a small cluster, not distance from the whole matrix, drives their centrality collapse — the concentrated-divergence mechanism Chapter 13 first described (§13.5's propositionbox) and Chapter 15 confirmed independently. West Bengal, Delhi NCT, and now Haryana sit at the opposite corner (0.871/0.161, 0.938/0.219, and 0.788/0.091): low isolation rank, high centrality, the

broadly-spread-closeness mechanism West Bengal's own §3.16 profile identifies and Chapter 18's own §18.4 names explicitly for Haryana, whose degree-rank/N figure, 0.091, is in fact the single most extreme value in either column of this table — the closest any tested pool has come to maximal weighted-degree centrality within its own system. Punjab alone sits near the table's own centre on both measures (0.586/0.448), which is consistent with Chapter 14's own finding that Punjab's addition strengthened τ through every other pool's figures shifting slightly, not through Punjab's own extreme placement — the only one of the six additions for which that mechanism, rather than the addition's own outlier position, is the one Chapter 14 itself named as responsible for the change in Table 21.1.

21.2 The Family-Excluded Deviation, Six Points in One Table

A second statistic follows the same pattern: six deviations of Indian sub-pool family-excluded isolation from whole-country India, each reported once, in its own chapter, against a progressively larger family-excluded universe.

Pool	Zone isol.	Diff. from India (0.2755)	Source
Andhra Pradesh	0.3432	+0.0677	Ch. 13, Table 13.4
Punjab	0.2950	+0.0195	Ch. 14, Table 14.4
Kerala	0.3445	+0.0690	Ch. 15, Table 15.4
West Bengal	0.2834	+0.0079	Ch. 16, Table 16.4
Delhi NCT	0.2662	−0.0093	Ch. 17, Table 17.4
Haryana	0.2982	+0.0227	Ch. 18, Table 18.3

Table 21.3: Family-excluded isolation of the six tested Indian sub-pools against whole-country India's own family-excluded figure (0.2755 bits, as recomputed at each successive stage of the family-excluded universe; Chapters 13–17 and 18 report this figure changing only in the fourth decimal place across the six stages). Every entry copied unchanged from its source table.

Ordered by absolute deviation rather than by chapter, the six figures run Kerala (0.0690) > Andhra Pradesh (0.0677) > Haryana (0.0227) > Punjab (0.0195) > West Bengal (0.0079) > Delhi NCT (0.0093, the sole negative value). This is the same two-group split Table 21.2 produced from an entirely different statistic, with Haryana's addition sharpening rather than blurring it: Kerala and Andhra Pradesh, the two high-isolation/low-centrality pools, still carry the two largest family-excluded deviations, both well over three times Haryana's own and Punjab's alike; Delhi NCT, the most centrally connected of the six tested before Haryana, remains the only pool less isolated from its family-excluded universe than whole-country India itself, and Haryana's own mid-range figure — larger than Punjab's, smaller than either Dravidian-zone outlier — falls where its broadly-spread-closeness degree-centrality profile (§18.4) would predict rather than at either extreme. Two independently computed statistics, over two different pool universes (the full matrix against the family-excluded one), assembled here for the first time side by side, agree on which two of the six Indian additions are the structurally unusual ones and which four are comparatively unremarkable in scale — a convergence Chapter 20's own closing discussion described in prose (§20's second section) without the table that makes it checkable at a glance.

21.3 What a Seventh Point Would, and Would Not, Add

Chapter 19 closed this volume's Indian sub-pool sequence at six tested instances and one identified, untested one — Haryana, admin1 `10`, resolved by name but never run through either statistic (§17.7) — on the stated ground that a scope decision, not a diminishing-returns judgement, was the honest reason to stop. The two tables assembled in this chapter do not overturn that decision; they sharpen the case for it by showing what it would concretely mean to run a seventh instance now that the trajectory, rather than any single addition, is visible in one place. A seventh τ value would extend Table 21.1 from seven points to eight; given that six of the existing seven already agree on direction, and that the association has already reached, at Delhi NCT, more than five times the two-tailed significance of the original twenty-six-pool figure ($p = 7.1 \times 10^{-7}$ against $p = 0.0030$), an eighth point's marginal contribution to the question "does this association hold up under further sub-national delimitation" is properly small — not zero, since Haryana's own isolation/degree placement is not knowable in advance, but small in the specific, now-demonstrable sense that no single addition among the last five has moved τ by more than 0.05 in either direction, and the question the statistic addresses has already accumulated six broadly consistent answers. A seventh family-excluded deviation would extend a five-point range that already spans $+0.0690$ to -0.0093 with two structurally distinct groups identified; a sixth point, wherever it fell inside or outside that range, would refine rather than reframe the two-group pattern Table 21.3 already shows.

The two paragraphs above were written before Chapter 18 existed, predicting what an eighth τ value and a sixth family-excluded deviation would look like without knowing either figure. Both predictions held. The eighth τ value, -0.5909 (Table 21.1), moved by -0.0103 from the seventh — smaller than the 0.05 threshold this section named as the boundary of a "properly small" marginal contribution, and consistent in direction with six of the preceding seven steps. The sixth family-excluded deviation, $+0.0227$ (Table 21.3), falls inside the five-point range this section predicted it would either fall inside or refine from outside — specifically between Punjab's $+0.0195$ and Andhra Pradesh's $+0.0677$, refining rather than reframing the two-group split exactly as predicted. Neither figure was known when this section's predictive claims were drafted, and neither required those claims to be revised once it was.

Chapter 19 distinguished an *open* item, whose resolution requires evidence external to this corpus, from an *extensible* one, whose method defines an unbounded further sequence with no internal stopping rule (§20, first section). This chapter's own contents belong to neither category exactly. Assembling Tables 21.1 and 21.3 required no new GeoNames extraction, no new pipeline run, and settled no open question; it is not extensible in Chapter 19's sense either, since there is no unbounded further sequence of comparable tables waiting to be built — exactly two statistics were reported enough times, in this volume, to be worth assembling this way, and both are now assembled. What this chapter offers is a third kind of closing work distinct from either: not computing a new instance, and not declining to, but reorganizing instances already computed and validated elsewhere into the single arrangement their own chapters' comparative language was already reaching for. It is offered as a legitimate, if narrower, alternative to extending the Indian sub-pool sequence a seventh time — one that adds no risk of the half-finished computation Chapter 19's own remark box on Haryana specifically declined to publish (§20), because nothing in it was ever run only partway.

Coda

Ten volumes and roughly 1,560 pages have applied two independently developed mathematical frameworks — one rooted in classical information theory, the other in graph theory, information geometry, and topological data analysis — to the same underlying question: what, if anything, is recoverable about historical linguistic substrates from the surface phonotactics of a modern, open, administratively-organised gazetteer. The two frameworks were built at different times, by different chapters' declared logic, and were never designed to be cross-checked against each other. That they nonetheless converge, independently and repeatedly, on the same handful of zones — Basque and Libya as structural outliers under every method that has been applied to them; the Arabic–Castilian pair as the most convergent instance of assimilation under both a bits-based and a graph-based reading — is the closest thing either series has produced to external validation of its own central premise. It is also, appropriately, not the last word: Chapter 7 above turns one strand of that convergence — isolation against centrality, on the real twenty-six-pool system — from a qualitative pattern into a measured one, $\tau = -0.4092$, $p = 0.0030$; extends the delimited-zone/whole-country comparison from a single documented case (Libya) to all twelve zones Networks Vol. III verified, finding a real association with one clearly documented exception (German-speaking Switzerland); adds, in Chapter 8, a twenty-seventh admin-verified pool found only after that chapter was drafted, on which the same isolation/ centrality association is not merely reproduced but strengthens (τ from -0.4092 to -0.4701), the same family-excluded isolation score of Item 2 places it within 0.004 bits of its own parent, closer than its sole Georgian sibling, and the new pool itself clusters, unremarkably and without resolving the question, inside the one grouping either series' own methodology already flags as genealogically open; and completes, in Chapter 9, the one comparison the earlier draft of this Coda had to leave open — the full thirty-three-zone Entropy-series graph, recomputed on its own original terms, over a directly measured shared-vocabulary divergence matrix rather than the original proxy formula, with the twenty-one previously undocumented zone definitions traced to their own source chapters rather than re-derived. The result is not a simple confirmation: Libya's global peripherality is independently reconfirmed, the Arabic–Castilian pair remains the closest edge within Iberia, but the single closest pair in the entire graph turns out to be Anatolian–Caucasian-TR, and Maghrebi centrality under the original proxy formula does not survive direct measurement. Ten volumes have now become eleven, and the balance sheet this Coda opened with is, for the first time, complete on its own terms.

One further item, smaller in scope than Chapters 7–9 but no less a matter of the same discipline, closes after this Coda's own balance sheet was first drawn: Chapter 10 runs Networks Vol. III's own named resampling protocol on Islamabad's 563-entry pool and finds, honestly, that its published figure does not exceed the noise floor twelve independent redraws establish — a negative result, reported as such, in the same spirit as Chapter 4's persistent-homology findings. Nouakchott's paired item does not close the same way: its 8-entry pool is smaller than the protocol's own $n = 40$ sample size, and no honest resampling check can be run on it at all. What Chapter 10 reports instead — a population-level divergence table and a leave-one-out jackknife, both explicitly flagged as artefacts of Laplace-prior dominance rather than toponymic signal — is offered as the more valuable kind of closure this volume has tried to practise throughout: not a result stretched past what eight real toponyms can support, but a precise account of why no stronger one is

currently available.

A second item closes in the same late window as Chapter 10, on the same discipline of applying a named protocol to a named target rather than inventing a substitute: Chapter 11 extends *Networks*, Vol. III, §20.2's Basque ranking-claim check — itself only a single-case proof of concept when originally run — to the three specific candidates its own roadmap named and left untried, Sindh, Corsica, and Sardegna. The result strengthens rather than merely repeats the Basque precedent: where Basque's isolation held in nine of thirteen independent draws, Sindh's and Corsica's hold in all thirteen, and Sardegna's in twelve of thirteen, its lone exception a near-tie of two hundredths of a radian. Four independent applications of the same construction, across two source volumes and this one, now agree that these zones' isolation claims are not artefacts of a single published seed.

Two further chapters close this expanded balance sheet on the same discipline. Chapter 12 pushes two of Networks Vol. III's admin1-resolution zones one level finer, finding Bolzano/South Tyrol and Trento genuinely distinct within Trentino-Alto Adige, and Occitanie's two 2016 predecessor regions genuinely distinct within it, while stating directly, rather than working around, why the third named candidate — Croatian Istria — has no admin2 substructure in `HR.txt` for the split to act on. Chapter 13 adds a twenty-eighth admin-verified pool, Andhra Pradesh, found by the same protocol Chapter 8 used for Kvemo Kartli and tested on the same two statistics: it obeys the same isolation/centrality sign every other pool has obeyed, but as the single most extreme case yet recorded — the system's new lowest-centrality pool despite only middling isolation, and the largest single deviation yet recorded on the family-excluded-isolation statistic, large enough to measurably loosen rather than strengthen that statistic's own cross-zone correlation. Both chapters extend the same open agenda items named below rather than closing them outright, and both say so explicitly rather than overstating what one further pool or one further admin2 split can establish.

A final chapter closes this balance sheet a second time over on the same discipline that opened it. Chapter 14 adds a twenty-ninth pool, Punjab, by the identical admin1-identification protocol Chapters 8 and 13 already established, and finds a genuinely different pattern from either predecessor: not an extreme isolation/centrality mismatch, as Andhra Pradesh was, but the closest cross-border pair the entire matrix contains — Punjab and whole-country Pakistan, at 0.0791 bits, the ninth-smallest divergence among all non-parent pairs in the system — and an isolation/degree association that strengthens substantially rather than merely persisting ($\tau = -0.5172$, up from -0.4709). Read together with the Turkey–Caucasus–India–Pakistan cluster's continued robustness to three independent subnational tests now (Chapters 8, 13, and 14), the balance sheet this Coda draws is, once again, not a simple confirmation: what a delimited zone reveals about the system it joins depends on which zone it is, and this volume has now shown three different shapes that revelation can take from the same six-country open cluster alone.

This balance sheet closed once more, three times over, in the chapters immediately following Punjab, on the same discipline throughout. Chapter 15 adds a thirtieth pool, Kerala, and records the system's first genuine exception to a pattern every prior delimited zone had obeyed without exception: its closest pool in the full matrix is not its own whole-country parent but Andhra Pradesh, its Dravidian-substrate sibling, marginally closer than India itself — together with the largest single family-excluded deviation yet recorded, narrowly exceeding Andhra Pradesh's own. Chapter 16 adds a thirty-first pool, West Bengal, the mirror case: the smallest deviation of the four sizeable Indian sub-pools tested to that point, and the system's most centrally connected Indian instance, closing the portion of the untested-admin1-units item that concerns pools of more than a thousand records. Chapter 17 closes the item's own named residue list in full, adding a thirty-second

pool, Delhi NCT, which repeats Kerala's non-parent-closest-neighbour pattern in a second, independent instance — its closest pool is West Bengal, not whole-country India — and adds a further first: the only negative family-excluded-isolation deviation among the five Indian sub-pools this volume has now tested. The same chapter also resolves, without adding it to any comparison, the one admin1 code Chapter 13's own residue table had left ambiguous: code `10` is Haryana, its apparent multi-state footprint traced to a single duplicated-population village record rather than to any genuine spread. Six concrete Indian instances are now on the record — five tested in full, one identified and explicitly held back from testing — and the item Networks, Vol. III, §29.5 first named is, on its own terms, exhausted rather than merely extended again.

Appendix A: Reproducibility — The Chapter 7 Pipeline

The three computations of Chapter 7 reuse, without modification, the extraction and divergence pipeline documented in *Toponomastic Networks*, Vol. III, §1.2 and Ch. 28, applied here to the raw GeoNames files supplied with this session. The code below is condensed for reference; every routine name matches the source volume's own Appendix A convention.

```
def normalize(name):
    nfd = unicodedata.normalize('NFD', name)
    stripped = ''.join(c for c in nfd
                       if unicodedata.category(c) != 'Mn')
    return re.sub(r'[^a-z]', '', stripped.lower())

def load_pool(country_code, admin1_set=None):
    # column 2 = asciiname, column 6 = feature class,
    # column 10 = admin1 code (0-indexed)
    names = []
    for parts in tsv_rows(f"{country_code}.txt"):
        if parts[6] not in ('P', 'T'):
            continue
        if admin1_set is not None and parts[10] not in admin1_set:
            continue
        az = normalize(parts[2])
        if len(az) >= 3:
            names.append(az)
    return names

LETTERS = list('abcdefghijklmnopqrstuvwxyz') + ['#']
IDX = {c: i for i, c in enumerate(LETTERS)}

def bigram_vector(name_az):
    s = '#' + name_az + '#'
    vec = np.zeros(27 * 27)
    for i in range(len(s) - 1):
        vec[IDX[s[i]] * 27 + IDX[s[i + 1]]] += 1
    return vec

def aggregate_dist(names, eps=1.0, K=729):
    total = np.full(K, eps)          # Laplace smoothing, eps=1
    for nm in names:
        total += bigram_vector(nm)
    return total / total.sum()

def js_divergence(p, q):
    m = 0.5 * (p + q)
    kl = lambda a, b: np.sum(np.where(a > 0, a * np.log2(a / b), 0.0))
    return 0.5 * kl(p, m) + 0.5 * kl(q, m)
```

```
# Item 1: isolation and affinity-weighted degree, 26-pool matrix
EPS_AFF = 0.05                        # Methodological Volume, Sec. 9.1
isolation = {i: mean(JS[i, j] for j in others)}
degree    = {i: sum(1.0 / (JS[i, j] + EPS_AFF) for j in others)}
tau, p = kendalltau(isolation_ranks, degree_ranks)

# Item 2: family-excluded isolation, 26+12-pool matrix
# (same aggregate_dist / js_divergence; a pool's own parent and
#  sibling sub-zones are excluded from its own isolation average)

# Item 4 (Chapter 8): 27th pool, Kvemo Kartli, admin1 68 in GE.txt
KVEMO_KARTLI = {'68'}                 # verified against 7 known towns
pool_27 = load_pool('GE', KVEMO_KARTLI)         # 365 names
# isolation/degree/tau recomputed on the full 27x27 matrix, not patched
# onto the published 26x26 one; Item 1's own formulas, unchanged

# Item 5 (Chapter 8, Sec. 8.5): 13th delimited zone, family-excluded
FAMILY = {'Libya': ['Libya','Tripolitania','Cyrenaica','Fezzan'],
          'Mauritania': ['Mauritania','Sahel','Sahara'],
          'Turkey': ['Turkey','Central Anatolia','Cappadocia core'],
          'Switzerland': ['Switzerland','German-speaking cantons'],
          'Georgia': ['Georgia','Samtskhe-Javakheti','Kvemo Kartli'],
          'Azerbaijan': ['Azerbaijan','Talysh core'],
          'United Kingdom': ['United Kingdom','Northern Ireland'],
          'Ireland': ['Ireland','Ulster rump']}
# family_excluded_isolation(pool, target) = mean JSD over every pool
# NOT in target's own family list (singleton family if not listed above)
```

Libya (Tripolitania: 77, 81, 68, 67, 80, 74, 82, 72, 76; Cyrenaica: 69, 66, 63, 70, 79, 83, 65; Fezzan: 75, 73, 71, 84, 78, 64 — Networks Vol. III, Table 25.1). Mauritania (Sahel: 01, 02, 03, 04, 05, 06, 10; Sahara: 07, 08, 09, 11, 12 — Table 26.1). Turkey (Central Anatolia: 68, 71, 38, 58, 66, 40, 50, 73, 75, 82, 78; Cappadocia core: 50, 73, 38, 75 — §§2.1, 3.1). Switzerland, German-speaking core (ZH, SG, UR, TG, LU, SZ, GL, NW, AG, OW, SO, BL, ZG, SH, AR, AI, BS — Table 4.1). Georgia (Samtskhe-Javakheti: 72 — §11.1). Azerbaijan (Talysh core: 29, 30, 08, 31, 32, 66 — §11.1). United Kingdom (Northern Ireland: NIR). Ireland (Ulster rump: U). Every code is copied from the named table or section of Networks Vol. III, not re-derived; none was guessed.

Georgia (Kvemo Kartli: 68), self-verified in the present session, not copied from any source volume, against the administrative-seat records of seven towns in `GE.txt` — Rustavi, Marneuli, Bolnisi, Gardabani, Dmanisi, Tsalka, Tetritsqaro — all sharing admin1=68 and none sharing Samtskhe-Javakheti's own code, 72.

Appendix B: Reproducibility — The Chapter 9 Thirty-Three-Zone Pipeline

Chapter 9's computation reuses the same `normalize`, `bigram_vector`, and `js_divergence` routines listed in Appendix A, applied to all thirty-three zones at once rather than to the twenty-six/twenty-seven-pool subset of Chapters 7–8. The zone table below is the complete definition set corresponding to Table 9.1; every code is copied from the named Entropy-series chapter, none guessed.

```
ZONES = {
  # Vol. II -- Iberia (Entropy II, SS.28-.29)
  "Basque":      [("ES", {"59","32"}, None),
                  ("FR", {"75"}, {"64","40"})],
  "Galician":    [("ES", {"34","58","39"}, None),
                  ("PT", {"04","20","05","21","06","22","11"}, None)],
  "Catalan":     [("ES", {"56","60","07"}, None)],
  "Arabic":      [("ES", {"51","31"}, None),
                  ("ES", {"54"}, {"AB","CR","TO"})],
  "Castilian":   [("ES", {"27","52","55","29","57"}, None),
                  ("ES", {"54"}, {"GU","CU"})],
  # Vol. III -- Central Europe (Entropy III, App. G.1 / Ch. 2)
  "Alpine-Celtic":   [("CH", None, None)],
  "Norican-Celtic":  [("AT", None, None)],
  "Illyrian-SI":     [("SI", None, None)],
  "Illyrian-HR":     [("HR", None, None)],
  "Venetian-Raetic": [("IT", {"20","17"}, None)],
  "Etruscan":        [("IT", {"16","18","07","05"}, None)],
  # Vol. IV -- British Isles (Entropy IV, Ch. 2.1)
  "Brittonic-Wales":            [("GB", {"WLS"}, None)],
  "PreCeltic-England":          [("GB", {"ENG"}, None)],
  "Goidelic-Ireland":           [("IE", None, None)],
  "Goidelic-NIreland":          [("GB", {"NIR"}, None)],
  "Pictish-Celtic-Scotland":    [("GB", {"SCT"}, None)],
  # Vol. V -- Indian Subcontinent (Entropy V, S2.3)
  "PK-Plains":     [("PK", {"04","05","02"}, None)],
  "PK-NW":         [("PK", {"03","06","07"}, None)],
  "Indo-Aryan-N":  [("IN", {"16","37","38","33","40",
                            "25","09","39","24"}, None)],
  "Dravidian":     [("IN", {"34","11","12","35","36"}, None)],
  "Munda":         [("IN", {"21"}, None)],
  "Indo-Aryan-E":  [("IN", {"19","03"}, None)],
  # Vol. VI -- Anatolia and Caucasus (Entropy VI, Tables 2.2-2.4, S2.2.4)
  "Azerbaijani":     [("AZ", None, None)],
  "Greek-Byzantine": [("TR", {"35","09","48","45","10",
                              "17","16","41","59"}, None)],
  "Anatolian":       [("TR", {"19","26","38","40","50","58",
                              "60","66","68","71","73","75","79"}, None)],
```

```
  "Caucasian-TR":    [("TR", {"04","08","21","25","63",
                              "65","76","84","86"}, None)],
  "Georgian":        [("GE", None, None)],
  "Armenian":        [("AM", None, None)],
  # Vol. VII -- Maghreb (Entropy VII, Preface/Ch. 1)
  "Mauritania": [("MR", None, None)],
  "Morocco":    [("MA", None, None)],
  "Tunisia":    [("TN", None, None)],
  "Algeria":    [("DZ", None, None)],
  "Libya":      [("LY", None, None)],
}

# pool_names(country, admin1_set, admin2_set) filters feature classes
# P/T, restricts by admin1 (and admin2 where given), normalizes and
# deduplicates; a zone with multiple (country, admin1, admin2) sources
# concatenates their pools before deduplication and sampling.
#
# Sampling: fixed seed 42, n=3000 per zone (full pool if smaller);
# aggregate_dist as in Appendix A (Laplace eps=1, K=729).
#
# JSD[i][j] = js_divergence(dist_i, dist_j) for all i != j, giving the
# full 33x33 symmetric matrix (528 distinct entries).
#
# Centrality (Methodological Volume S9.1 formulas, unchanged):
EPS_AFF = 0.05
# degree[i]  = sum(1.0 / (JSD[i][j] + EPS_AFF) for j != i)
# eigenvector centrality: networkx.eigenvector_centrality_numpy on the
#   complete graph with edge weight 1/(JSD[i][j] + EPS_AFF)
# betweenness: networkx.betweenness_centrality(weight='JSD') on the
#   k-nearest-neighbour sparsification of the JSD distance matrix;
#   k found by minimal search over k=1,2,3,... until connected
#   (k=3: disconnected; k=4: connected -- matches Networks Ch. 9)
```

Appendix C: Reproducibility — The Chapter 10 Resampling Pipeline

Chapter 10 reuses `normalize`, `load_pool`, `bigram_vector`, `aggregate_dist`, and `js_divergence` from Appendix A verbatim, and the Fisher–Rao sampling, persistence, and bottleneck-distance routines Networks Vol. III's own Appendix 12/13/19 already document (`ripser`, `persim.bottleneck`) — none of this is repeated below. The only code genuinely new to this chapter is the leave-one-out jackknife loop, since neither source volume's appendices contain one.

```
# Nouakchott: leave-one-out jackknife, JSD against Mauritania whole
# (aggregate_dist, js_divergence: Appendix A, unchanged)
jackknife = []
for i in range(len(nouakchott_names)):
    sub = nouakchott_names[:i] + nouakchott_names[i+1:]
    j = js_divergence(aggregate_dist(sub), aggregate_dist(mr_whole))
    jackknife.append(j)
```

Every figure in Chapter 10 was produced by Appendix A's unchanged routines plus the resampling and persistence code already documented in the cited Networks Vol. III appendices, applied directly to PK.txt and MR.txt, with only the jackknife loop above added new. No figure in Chapter 10 is estimated, interpolated, or carried over from either source volume without independent recomputation here.

Appendix D: Reproducibility — The Chapter 11 Ranking-Claim Pipeline

Chapter 11 reuses `normalize`, `load_pool`, and `bigram_vector` from Appendix A verbatim, and the `sample_and_fr_matrix`, `frechet_mean_sphere`, and `pairwise_frechet_distances` routines *Networks*, Vol. III's own Appendix 13 documents (§13, items 4 and 6–7 there), also unchanged — none of this is repeated below. The only code genuinely new to this chapter is the per-draw bookkeeping that runs all pools of a given case (Pakistan's eight, France's four, or Italy's six) at a shared seed and records whether the named pool supplied the overall maximum of that draw's distance matrix.

```
# Chapter 11: per-case ranking-claim resampling check
# (sample_and_fr_matrix, frechet_mean_sphere,
#  pairwise_frechet_distances: Networks Vol. III App. 13, unchanged)
SEEDS = [0, 1000, 2000, 3000, 4000, 5000, 6000,
         7000, 8000, 9000, 10000, 11000]

def one_draw(pool_codes, pool_loader, seed):
    means = {}
    for lab in pool_codes:
        _, sq = sample_and_fr_matrix(pool_loader[lab], n=40, seed=seed)
        means[lab] = frechet_mean_sphere(sq)
    D = pairwise_frechet_distances(means, pool_codes)
    return D

def ranking_check(pool_codes, pool_loader, target):
    def result(seed):
        D = one_draw(pool_codes, pool_loader, seed)
        i_t = pool_codes.index(target)
        target_max = max(D[i_t, j] for j in range(len(pool_codes))
                         if j != i_t)
        overall_max = max(D[i, j] for i in range(len(pool_codes))
                          for j in range(len(pool_codes)) if i != j)
        is_source = any(abs(D[i_t, j] - overall_max) < 1e-12
                        for j in range(len(pool_codes)) if j != i_t)
        return target_max, overall_max, is_source
    original = result(42)
    resamples = [result(s) for s in SEEDS]
    return original, resamples

# Pool sets, admin1 codes copied from Networks Vol. III Tables 16.1,
# 17.1, 19.1 -- none re-derived:
PK_POOLS = {"Balochistan": "02", "KP": "03", "Punjab": "04",
            "Sindh": "05", "AJK": "06", "GB": "07", "Islamabad": "08"}
            # + whole-country (admin1_set=None)
FR_POOLS = {"Brittany": "53", "Corsica": "94", "Occitanie": "76"}
            # + whole-country
```

```
IT_POOLS = {"Trentino": "17", "ValleAosta": "19", "FriuliVG": "06",
            "Sardegna": "14", "Veneto": "20"}
            # + whole-country
```

Every figure in Chapter 11 was produced by Appendix A's unchanged extraction routines plus the unmodified Fréchet-mean/Fisher–Rao code Networks Vol. III's own Appendix 13 already documents, applied directly to PK.txt, FR.txt, and IT.txt, with only the source-identification bookkeeping above added new. Every pool's n and every seed-42 headline distance was checked against its published source-volume figure (§11.1) before any resampling draw was run.

Appendix E: Reproducibility — The Chapter 12 Admin2-Refinement Pipeline

Chapter 12 reuses `normalize`, `bigram_vector`, `aggregate_entropy`, `js_divergence`, `sample_and_fr_matrix`, and `frechet_mean_sphere` from Appendices A and D verbatim, and `ripser` exactly as *Networks*, Vol. III's own Appendix 19 applies it — none of this is repeated below. The only code genuinely new to this chapter is `load_pool_admin2`, which restricts `load_pool`'s existing single-field filter to a joint admin1/admin2 condition, and the department-code sets used to split Occitanie.

```
# Chapter 12: admin2-restricted pool loader
# (normalize, aggregate_entropy, js_divergence, sample_and_fr_matrix,
#  frechet_mean_sphere: Appendices A and D, unchanged)
def load_pool_admin2(country_code, admin1, admin2_set):
    names = []
    for parts in tsv_rows(f"{country_code}.txt"):
        if parts[6] not in ('P', 'T'):
            continue
        if parts[10] != admin1:
            continue
        if parts[11] not in admin2_set:
            continue
        az = normalize(parts[2])
        if len(az) >= 3:
            names.append(az)
    return names

# Trentino-Alto Adige (admin1=17 in IT.txt): self-reported admin2
# letter codes, no external table needed.
BOLZANO = {"BZ"}
TRENTO  = {"TN"}

# Occitanie (admin1=76 in FR.txt): department numbers, split between
# the two 2016 predecessor regions; department-to-region assignment
# checked against public administrative sources (Wikipedia; Britannica;
# Region Occitanie's own site), not recalled from memory alone.
MIDI_PYRENEES = {"09","12","31","32","46","65","81","82"}
LANGUEDOC_ROUSSILLON = {"11","30","34","48","66"}
```

Every figure in Chapter 12 was produced by Appendices A and D's unchanged routines plus `ripser`, applied directly to IT.txt and FR.txt, with only the two-field `load_pool_admin2` filter above added new. Both whole-region rows (Tables 12.1 and 12.2) were checked against their published source figures before either split was trusted. Croatian Istria's admin2 refinement was not attempted at all, for the reason §12.3 states directly; no code was written toward it.

Appendix F: Reproducibility — The Chapter 13 Andhra Pradesh Pipeline

Chapter 13 reuses, unchanged, `normalize`, `load_pool`, `bigram_vector`, `aggregate_dist`, and `js_divergence` from Appendix A, and the family-excluded isolation routine implicit in Item 2 there. The only new inputs are the admin1 code identified below and its addition to the pool and family dictionaries already defined in Appendices A and B.

```
# Chapter 13: 28th pool, Andhra Pradesh, admin1 02 in IN.txt
VERIFY_TOWNS = ["Visakhapatnam","Vijayawada","Guntur","Nellore",
                "Kurnool","Kadapa","Tirupati","Anantapur","Chittoor"]
# Each located in IN.txt by asciiname (highest-population record kept
# where a name recurs); all nine share admin1 == '02'.

ANDHRA_PRADESH = {'02'}
pool_28 = load_pool('IN', ANDHRA_PRADESH)      # 29,029 names

# Re-validation: the 27-pool system (POOLS_26 + Kvemo Kartli) is rebuilt
# and its isolation/degree/tau figures checked against Table 8.2 (tab:item1-27)
# before Andhra Pradesh is added to any comparison.

# Isolation/degree/tau recomputed on the full 28x28 matrix, exactly as
# Chapter 8 did for the 27x27 matrix -- not patched onto the published one.

# Family-excluded isolation: India has no other documented delimited
# sub-zone, so Andhra Pradesh's own family (for exclusion purposes) is
# the singleton set {India, Andhra Pradesh}:
FAMILY['India'] = ['India', 'Andhra Pradesh']

# Remaining untested admin1 codes, for the record (not computed further):
# Punjab '23' (11,508), Kerala '13' (7,153), West Bengal '28' (4,309),
# Delhi NCT '07' (617); admin1 '10' (7,246) excluded from this list as
# unverified -- its top-population entries span more than one state.
```

India (Andhra Pradesh: `02`), self-verified in the present session against nine towns in `IN.txt` — Visakhapatnam, Vijayawada, Guntur, Nellore, Kurnool, Kadapa, Tirupati, Anantapur, Chittoor — all sharing admin1=`02`, distinct from Telangana's own code, `40`, already in use in Entropy Vol. V's Indo-Aryan-N pool (Appendix B above).

Every figure in Chapter 13 was produced by Appendix A's unchanged `load_pool/aggregate_dist/js_divergence` routines, applied directly to `IN.txt` (already loaded for the whole-country India pool of Chapter 7) with the single new admin1 filter `{'02'}`. No new extraction or divergence code was written; the only genuinely new artefact is the admin1 identification of §13.1 itself.

Appendix G: Reproducibility — The Chapter 14 Punjab Pipeline

Chapter 14 reuses, unchanged, `normalize`, `load_pool`, `bigram_vector`, `aggregate_dist`, and `js_divergence` from Appendix A, and the family-excluded isolation routine and UPGMA and cophenetic-correlation procedure already used in Chapter 13. The only new inputs are the admin1 code identified below and its addition to the pool and family dictionaries already defined in Appendices A and B and extended in Chapter 13.

```
# Chapter 14: 29th pool, Punjab, admin1 23 in IN.txt
VERIFY_TOWNS = ["Amritsar","Ludhiana","Jalandhar","Patiala","Bathinda"]
# Each located in IN.txt by asciiname (highest-population record kept
# where a name recurs); all five share admin1 == '23'.

PUNJAB = {'23'}
pool_29 = load_pool('IN', PUNJAB)      # 11,508 names

# Re-validation: the 28-pool system (POOLS_26 + Kvemo Kartli + Andhra
# Pradesh) is rebuilt and its isolation/degree/tau figures checked
# against Table 13.2 (tab:item1-28) before Punjab is added to any
# comparison.

# Isolation/degree/tau recomputed on the full 29x29 matrix, exactly as
# Chapters 8 and 13 did for their own extensions -- not patched onto
# either published matrix.

# Family-excluded isolation: Punjab joins Andhra Pradesh inside India's
# own family (for exclusion purposes):
FAMILY['India'] = ['India', 'Andhra Pradesh', 'Punjab']

# Cross-border scan (Section 14.3): every pairwise JSD in the 29x29
# matrix excluding the eight documented parent/child pairs, sorted
# ascending, to place Punjab-Pakistan's rank honestly rather than
# asserting it without a denominator.

# Remaining sizeable untested admin1 codes, for the record:
# Kerala '13' (7,153), West Bengal '28' (4,309), Delhi NCT '07' (617);
# admin1 '10' (7,246) still excluded as unverified -- its top-population
# entries span more than one state.
```

India (Punjab: 23), self-verified in the present session against five towns in `IN.txt` — Amritsar, Ludhiana, Jalandhar, Patiala, Bathinda — all sharing admin1=23, distinct from every other Indian admin1 code already in use elsewhere in this volume or in Entropy Vol. V's own Indian-Subcontinent zones (Appendix B above).

Every figure in Chapter 14 was produced by Appendix A's unchanged `load_pool/aggregate_dist/js_divergence` routines, applied directly to `IN.txt` (already loaded for the whole-country India pool of Chapter 7 and the Andhra Pradesh pool of Chapter 13) with the single new admin1 filter `{'23'}`, and by the same UPGMA/cophenetic procedure Chapter 13 already used. No new extraction or divergence code was written; the only genuinely new artefact is the admin1 identification of §14.1 itself and the cross-border ranking scan of §14.3.

Appendix H: Reproducibility — The Chapter 15 Kerala Pipeline

Chapter 15 reuses, unchanged, `normalize`, `load_pool`, `bigram_vector`, `aggregate_dist`, and `js_divergence` from Appendix A, and the family-excluded isolation routine and UPGMA and cophenetic-correlation procedure already used in Chapters 13 and 14. The only new inputs are the admin1 code identified below and its addition to the pool and family dictionaries already defined in Appendices A and B and extended in Chapters 13 and 14.

```
# Chapter 15: 30th pool, Kerala, admin1 13 in IN.txt
VERIFY_TOWNS = ["Thiruvananthapuram", "Kochi", "Kozhikode",
                "Thrissur", "Kollam", "Kannur"]
# Each located in IN.txt by asciiname, restricted to feature class 'P'
# and the highest-population record where a name recurs; all six share
# admin1 == '13'.

KERALA = {'13'}
pool_30 = load_pool('IN', KERALA)      # 7,153 names

# Re-validation: the 29-pool system (POOLS_26 + Kvemo Kartli + Andhra
# Pradesh + Punjab) is rebuilt and its isolation/degree/tau figures
# checked against Table 14.1 (tab:item1-29) before Kerala is added to
# any comparison.

# Isolation/degree/tau recomputed on the full 30x30 matrix, exactly as
# Chapters 8, 13, and 14 did for their own extensions -- not patched
# onto any previously published matrix.

# Family-excluded isolation: Kerala joins Andhra Pradesh and Punjab
# inside India's own family (for exclusion purposes):
FAMILY['India'] = ['India', 'Andhra Pradesh', 'Punjab', 'Kerala']

# Closest-pool scan (Section 15.3): Kerala's own row of the 30x30
# matrix, sorted ascending, to check directly whether its own parent
# (India) is still its single closest pool -- it is not: Andhra
# Pradesh, at 0.1225 bits, is closer than India, at 0.1345 bits.

# Sibling-pair check (Section 15.3, Proposition): the same matrix's
# Andhra-Pradesh/Kerala cell, 0.1225 bits, compared directly against
# Andhra Pradesh's own divergence from whole-country India, 0.1358
# bits (Table 13.2, tab:andhra-extremes) -- both read from the one
# real matrix, not estimated or interpolated.

# Remaining untested admin1 codes, for the record (not computed further):
# West Bengal '28' (4,309), Delhi NCT '07' (617); admin1 '10' (7,246)
# still excluded from this list as unverified -- its top-population
# entries span more than one state.
```

India (Kerala: `13`), self-verified in the present session against six towns in `IN.txt` — Thiruvananthapuram, Kochi, Kozhikode, Thrissur, Kollam, Kannur — all sharing admin1=`13`, distinct from Andhra Pradesh's own code, `02`, Punjab's own code, `23`, and every other Indian admin1 code already in use elsewhere in this volume or in Entropy Vol. V's own Indian-Subcontinent zones (Appendix B above).

Every figure in Chapter 15 was produced by Appendix A's unchanged `load_pool/aggregate_dist/js_divergence` routines, applied directly to `IN.txt` (already loaded for the whole-country India pool of Chapter 7 and the Andhra Pradesh and Punjab pools of Chapters 13 and 14) with the single new admin1 filter `{'13'}`, and by the same UPGMA/cophenetic procedure Chapters 13 and 14 already used. No new extraction or divergence code was written; the only genuinely new artefacts are the admin1 identification of §15.1 itself and the non-parent-closest-neighbour finding of §15.3, which required no additional computation beyond reading the already-assembled 30×30 matrix's own Kerala row.

Appendix I: Reproducibility — The Chapter 16 West Bengal Pipeline

Chapter 16 reuses, unchanged, `normalize`, `load_pool`, `bigram_vector`, `aggregate_dist`, and `js_divergence` from Appendix A, and the family-excluded isolation routine and UPGMA and cophenetic-correlation procedure already used in Chapters 13–15. This appendix was not included with the chapter at the time of drafting; it is added here, retroactively, for consistency with Appendices F–H's own treatment of Chapters 13–15, and describes no computation beyond what Chapter 16 itself already reports.

```
# Chapter 16: 31st pool, West Bengal, admin1 28 in IN.txt
VERIFY_TOWNS = ["Kolkata", "Howrah", "Durgapur", "Siliguri",
                "Asansol", "Malda"]
# Each located in IN.txt by asciiname, restricted to feature class 'P'
# and the highest-population record where a name recurs; all six share
# admin1 == '28'. Asansol's own highest-population record carries the
# diacritic asciiname "Aasansol"; two zero-population admin1=38 (Bihar)
# duplicates share the plain spelling but are excluded by the
# highest-population rule.

WEST_BENGAL = {'28'}
pool_31 = load_pool('IN', WEST_BENGAL)      # 4,309 names

# Re-validation: the 30-pool system (POOLS_26 + Kvemo Kartli + Andhra
# Pradesh + Punjab + Kerala) is rebuilt and its isolation/degree/tau
# figures checked against Table 15.1 (tab:item1-30) before West Bengal
# is added to any comparison.

# Isolation/degree/tau recomputed on the full 31x31 matrix, exactly as
# Chapters 8, 13, 14, and 15 did for their own extensions.

# Family-excluded isolation: West Bengal joins Andhra Pradesh, Punjab,
# and Kerala inside India's own family (for exclusion purposes):
FAMILY['India'] = ['India', 'Andhra Pradesh', 'Punjab', 'Kerala',
                   'West Bengal']

# Closest-pool scan (Section 16.1): West Bengal's own row of the 31x31
# matrix, sorted ascending -- India remains its closest pool at 0.0377
# bits, unlike Kerala's own non-parent-closest result.

# Remaining untested admin1 codes, for the record: Delhi NCT ('07',
# 617 records) and the still-unverified code '10' (7,246 records,
# spanning more than one state on a naive scan -- see Chapter 17,
# Section 17.7, for the resolution).
```

India (West Bengal: `28`), self-verified in the present session against six towns in `IN.txt` — Kolkata, Howrah, Durgapur, Siliguri, Asansol, Malda — all sharing admin1=`28`, distinct from Andhra Pradesh's `02`, Punjab's `23`, and Kerala's `13`.

Every figure in Chapter 16 was produced by Appendix A's unchanged `load_pool/aggregate_dist/js_divergence` routines, applied directly to `IN.txt` with the single new admin1 filter `{'28'}`, and by the same UPGMA/cophenetic procedure Chapters 13–15 already used. No new extraction or divergence code was written.

Appendix J: Reproducibility — The Chapter 17 Delhi NCT Pipeline

Chapter 17 reuses, unchanged, `normalize`, `load_pool`, `bigram_vector`, `aggregate_dist`, and `js_divergence` from Appendix A, and the family-excluded isolation routine and UPGMA and cophenetic-correlation procedure already used in Chapters 13–16. The only new inputs are the admin1 code identified below, its addition to the pool and family dictionaries already defined in Appendices A and B and extended in Chapters 13–16, and the separate admin1-code cross-reference of §17.7, which uses the identical `load_pool`/highest-population-selection routine but builds no new distance matrix.

```
# Chapter 17: 32nd pool, Delhi NCT, admin1 07 in IN.txt
VERIFY_TOWNS = ["Delhi", "Najafgarh", "Narela", "Rohini", "New Delhi"]
# Each located in IN.txt by asciiname, restricted to feature class 'P'
# and the highest-population record where a name recurs; all five share
# admin1 == '07'. A sixth candidate, Dwarka, was tested and rejected:
# its highest-population P-class record nationally is Gujarat's Dwarka
# (admin1=09, pop 38,873), not Delhi's own zero-population Dwarka
# record (admin1=07) -- a same-name collision, not a coding error.

DELHI_NCT = {'07'}
pool_32 = load_pool('IN', DELHI_NCT)        # 617 names

# Re-validation: the 31-pool system (POOLS_26 + Kvemo Kartli + Andhra
# Pradesh + Punjab + Kerala + West Bengal) is rebuilt and its
# isolation/degree/tau figures checked against Table 16.1
# (tab:item1-31) before Delhi NCT is added to any comparison.

# Isolation/degree/tau recomputed on the full 32x32 matrix, exactly as
# Chapters 8, 13, 14, 15, and 16 did for their own extensions -- not
# patched onto any previously published matrix.

# Family-excluded isolation: Delhi NCT joins Andhra Pradesh, Punjab,
# Kerala, and West Bengal inside India's own family (for exclusion
# purposes):
FAMILY['India'] = ['India', 'Andhra Pradesh', 'Punjab', 'Kerala',
                   'West Bengal', 'Delhi NCT']
# Full family-excluded universe rebuilt from raw GeoNames data: the
# 26 base pools, the 7 further Networks Vol. III zones of Appendix A
# (Central Anatolia, Cappadocia core, German-speaking Swiss cantons,
# Samtskhe-Javakheti, Talysh core, Northern Ireland, Ulster rump), and
# the 5 earlier Indian sub-pools -- 38 pools before Delhi NCT, 39 with
# Kvemo Kartli/Andhra/Punjab/Kerala/West Bengal already counted in the
# 38, matching Chapter 16's own stated "thirty-nine-pool" universe
# (thirty-eight plus West Bengal) exactly; Delhi NCT is the fortieth.

# Closest-pool scan (Section 17.3): Delhi NCT's own row of the 32x32
# matrix, sorted ascending -- West Bengal, at 0.0688 bits, is closer
# than whole-country India, at 0.0858 bits, the second non-parent-
```

```
# closest-neighbour case in this system after Kerala (Chapter 15).

# Admin1 code 10 resolution (Section 17.7, not part of the 32x32
# matrix): eight independently-known Haryana towns -- Faridabad,
# Gurugram, Rohtak, Hisar, Panipat, Karnal, Ambala, Sirsa -- located by
# the identical highest-population protocol, all resolving to
# admin1='10'. The single Gorakhpur record at admin1='10' whose
# population field (1,324,570) duplicates Uttar Pradesh's own
# same-named city exactly is documented as the source of Chapter 13's
# original "spans more than one state" caution.
```

India (Delhi NCT: `07`), self-verified in the present session against five towns in `IN.txt` — Delhi, Najafgarh, Narela, Rohini, New Delhi — all sharing admin1=`07`, distinct from Andhra Pradesh's `02`, Punjab's `23`, Kerala's `13`, and West Bengal's `28`. Admin1 code `10` (Haryana) is separately verified against eight towns, listed above, but its pool is not added to any distance matrix in this chapter.

The claim just above is accurate as a description of Chapter 17 itself, which is what this appendix documents, and is left unedited for that reason. It is no longer accurate as a description of this volume overall: Chapter 18, added in a later session, does add Haryana's pool to a distance matrix — the full 33×33 system — and Appendix K below documents that chapter's own pipeline on the same terms this appendix documents Chapter 17's. A reader relying on this remarkbox alone, out of context, would reach a conclusion this volume no longer holds; Appendix K is the place that conclusion is corrected.

Every figure in Chapter 17's main text was produced by Appendix A's unchanged `load_pool`/`aggregate_dist`/`js_divergence` routines, applied directly to `IN.txt` (already loaded for the whole-country India pool of Chapter 7 and the four earlier Indian sub-pools of Chapters 13–16) with the single new admin1 filter `{'07'}`, and by the same UPGMA/cophenetic procedure Chapters 13–16 already used. The family-excluded isolation table of §17.5 required rebuilding the full forty-pool universe (the twenty-six base pools, the seven further Networks Vol. III zones, and the six Indian sub-pools including Delhi NCT) from raw GeoNames files directly, since no prior chapter's own session state persists between chapters; this is the same rebuild-from-source discipline Chapters 8 and 13–16 already followed for their own family-excluded tables. The admin1=`10` identification of §17.7 used the identical `load_pool` routine restricted to name lookups only, with no aggregate distribution or divergence computed for that pool. No new extraction or divergence code was written anywhere in this chapter.

Appendix K: Reproducibility — The Chapter 20 Haryana Pipeline

Chapter 18 reuses, unchanged, `normalize`, `load_pool`, `bigram_vector`, `aggregate_dist`, and `js_divergence` from Appendix A, and the family-excluded isolation routine and UPGMA and cophenetic-correlation procedure already used in Chapters 13–17. The only new input is the admin1 code already identified in §17.7 and Appendix J, applied here for the first time to build a pool and enter it into a distance matrix.

```
# Chapter 20: 33rd pool, Haryana, admin1 10 in IN.txt
HARYANA = {'10'}
pool_33 = load_pool('IN', HARYANA)          # 7,245 names

# Re-validation: the 32-pool system (POOLS_26 + Kvemo Kartli + Andhra
# Pradesh + Punjab + Kerala + West Bengal + Delhi NCT) is rebuilt from
# raw GeoNames files and its isolation/degree/tau figures checked
# against Table 17.1 (tab:item1-32) before Haryana is added to any
# comparison. Kendall's tau on the reproduced 32-pool matrix:
# tau = -0.5806, p = 7.1e-07 (matches Chapter 17 digit-for-digit).

# Isolation/degree/tau recomputed on the full 33x33 matrix, exactly as
# Chapters 8 and 13-17 did for their own extensions -- not patched onto
# the published 32x32 matrix. UPGMA (average linkage) and cophenetic
# correlation likewise recomputed fresh on the 33x33 condensed distance
# matrix: cophenetic correlation = 0.8577 (32-pool figure: 0.8562).

# Family-excluded isolation: Haryana joins Andhra Pradesh, Punjab,
# Kerala, West Bengal, and Delhi NCT inside India's own family (for
# exclusion purposes):
FAMILY['India'] = ['India', 'Andhra Pradesh', 'Punjab', 'Kerala',
                   'West Bengal', 'Delhi NCT', 'Haryana']
# Family-excluded universe rebuilt from raw GeoNames data: the 26 base
# pools, the 7 further Networks Vol. III zones of Appendix A, Kvemo
# Kartli, and the six Indian sub-pools (Andhra Pradesh, Punjab, Kerala,
# West Bengal, Delhi NCT, Haryana) -- 40 pools total once Haryana is
# included, since Haryana both enters the universe being averaged over
# and is simultaneously excluded from its own family's averaging set,
# the same construction Chapters 8 and 13-17 already used for their own
# additions.

# Closest-pool scan (Section 20.3): Haryana's own row of the 33x33
# matrix, sorted ascending -- Punjab, at 0.0254 bits, is closer than
# whole-country India, at 0.0428 bits, the closest pair recorded
# anywhere in this volume and the third non-parent-closest-neighbour
# case after Kerala (Chapter 15) and Delhi NCT (Chapter 17).
```

Every figure in Chapter 18's main text was produced by Appendix A's unchanged `load_pool/aggregate_dist/ js_divergence` routines, applied directly to `IN.txt` (already loaded for the whole-country India pool of Chapter 7 and the six earlier Indian sub-pools of Chapters 13–17) with the single new admin1 filter `{'10'}`, and by the same UPGMA/cophenetic procedure Chapters 13–17 already used. Isolation, weighted degree, and the Kendall τ figures were checked against the reproduced 32×32 system before Haryana was added, exactly as every prior extension in this volume was validated. The family-excluded isolation table of §18.5 required rebuilding the full family-excluded universe from raw GeoNames files directly, following the same rebuild-from-source discipline as Appendices F–J. No new extraction or divergence code was written anywhere in this chapter; the single new computation this appendix documents is the application of already-published routines to a pool identified, but not previously tested, in Chapter 17.

A first pass of Table 18.2 carried the Degree rank column over incorrectly – the isolation and weighted-degree values were correct on first computation, but the ranks derived from the weighted-degree column were not freshly resorted after Haryana's row was inserted. Re-running the ranking directly from the (correct) weighted-degree column produced a revised Degree rank column, printed in the corrected Table 18.2 above; the error and its correction are recorded here rather than silently repaired, following this volume's own stated preference throughout for flagging a discrepancy explicitly over smoothing it away. No other statistic in Chapter 18 – the D_{JS} values of Table 18.1, the isolation values and Kendall τ of §18.4, the family-excluded isolation values of Table 18.3, or the cophenetic correlations of either the 32×32 or 33×33 system – required correction; each was independently re-verified against a direct recomputation from the raw GeoNames files before this volume's present text was finalised.

Bibliography

This volume's citations are almost entirely internal: every reference to "*Entropy*, Vol. N" or "*Networks*, Vol. N, §M" throughout the text points to a specific volume and section of this same author's own ten-volume series, named in full at each point of use rather than collected here. What follows instead are the two items External to that series that are named in the text (§19, Open Research Agenda), plus the primary data source underlying every figure in this volume and its nine predecessors.

- Villar, Francisco. *Estudios de celtibérico y de toponimia prerromana.* Acta Salmanticensia, Estudios Filológicos 260. Salamanca: Ediciones Universidad de Salamanca, 1995.
- Hoz, Javier de. *Historia lingüística de la Península Ibérica en la Antigüedad. II: El mundo ibérico prerromano y la indoeuropeización.* Manuales y Anejos de *Emerita* LI. Madrid: Consejo Superior de Investigaciones Científicas, 2011.
- GeoNames geographical database. `https://www.geonames.org/`. Data released under a Creative Commons Attribution 4.0 licence. All per-country extracts used in this volume and its two predecessor series (`IN.txt`, `TR.txt`, `CH.txt`, `GE.txt`, `AZ.txt`, `LY.txt`, `MR.txt`, `GB.txt`, `IE.txt`, and others named at point of use) were obtained from this source.

Villar (1995) and de Hoz (2011) are named once each, in the Open Research Agenda's "fine etymological validation" item, as sources for an etymologically-annotated Celtic sub-corpus that this volume's own computations do not use directly; the item remains proposed, not carried out, in every chapter above. No other named scholarly work appears anywhere in this volume's ten-volume source series or in this synthesis. A longer bibliography was considered and rejected: inventing further entries to make this list resemble a conventional monograph's apparatus would misrepresent a series whose evidentiary base is, throughout, a single open gazetteer and the two frameworks' own prior volumes, not a secondary literature this text has actually engaged with.